The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Number 1 March 1953 Volume 9

Material for *Biometrics* should be addressed to Miss Gertrude Cox, Institute of Statistics, Box 5457, Raleigh, North Carolina, except that authors residing in one of the following organized regions can expedite the handling of their papers by submitting them to the Assistant Editor for that region.

British Region: Dr. M. J. R. Healy, Rothamsted Exp. Sta., Harpenden, Herts, England (serving in Dr. Finney's absence); **Australasian Region:** Dr. E. A. Cornish, University of Adelaide, Adelaide, Australia; **French Region:** Dr. Georges Teissier, Faculte des Sciences de Paris, 1 rue V. Cousin, Paris, France

Material for Queries should go to Professor G. W. Snedecor, Statistical Laboratory, Iowa State College, Ames, Iowa.

Articles to be considered for publication should be submitted in triplicate.

The Biometric Society is an international society devoted to the mathematical and statistical aspects of biology and welcomes to membership biologists, mathematicians, statisticians and others who are interested in its objectives. Through its regional organizations the Society sponsors regional and local meetings. National secretaries serve the interest of members in Italy, Denmark, the Netherlands, India, Germany and Japan and there are many members "at large". Dues in the Society for 1953 for residents of the Western Hemisphere are as follows: Full membership including subscription to *Biometrics* is $7.00. Members of the Biometrics Section of the American Statistical Association who subscribe to the journal through that organization may become members of The Biometric Society on the payment of $3.00 annual dues. For members in other parts of the world, full membership including subscription to *Biometrics* is $4.50, except that members who subscribe to the journal through the American Statistical Association pay annual dues of $1.75. Information concerning the Society can be obtained from the Secretary, The Biometric Society, Drawer 1106, New Haven 4, Connecticut, U.S.A.

Annual subscription rates to non-members are as follows: For American Statistical Association Members, $4.00; for subscribers, non-members of either American Statistical Association or The Biometric Society, $7.00. Subscriptions should be sent to the Managing Editor, *Biometrics*, P. O. Box 5457, Raleigh, North Carolina, U.S.A.

Entered as second-class matter at the Post Office at New Haven, Conn., under the Act of March 3, 1879. Additional entry at Richmond, Va. Business Office, 52 Hillhouse Ave., New Haven, Conn. *Biometrics* is published quarterly—in March, June, September and December.

LABORATORY FLAVOR SCORING: TWO EXPERIMENTS IN INCOMPLETE BLOCKS[1]

J. W. Hopkins

Division of Applied Biology,
National Research Council,
Ottawa, Canada

ABSTRACT

Bitterness and saltiness of 9 test beverages comprising all combinations of 3 levels of two factors were scored on intensity scales by two groups of 24 unselected and untrained subjects. Group average scores (interpretable as estimates of corresponding population means) were sensibly linearly related to dosage of flavor additives. Differentiation between intensities of bitterness was slightly reduced at the highest level of saltiness. Replicate scores recorded for the various test samples were distributed with unequal variance and pronounced non-normality, but the latter was much reduced in individuals' score contrasts corresponding to single degrees of freedom. Separate analyses of variance of these contrasts were indicative of characteristic differences in individuals' scores analogous to complete block × treatment interactions, presumably related to sensory thresholds and acuities. Partition of the complete set of test samples into separately appraised balanced or unbalanced incomplete sets did not demonstrably modify either the mean or variance of recorded score contrasts.

INTRODUCTION

Subjective appraisals of flavor or palatability quantified by ranking or scoring are now widely used in food research and technology (1). Differing techniques have however been adopted, and more knowledge of the merits and limitations of each is desirable. For example, simultaneous taste stimuli may interact, resulting in variations in intensity of one stimulus modifying appraisals of another (8). Analogous interactions of stimuli observed concurrently, if operative when the samples constituting a test series must be appraised in incomplete sets, might result in the rating of an individual sample being biassed by the characteristics of the other samples occurring in the same set. The two

[1]N.R.C. Paper No. 2951.

experiments now reported were accordingly undertaken to study the extent, if any, of such biasses in flavor scores recorded in the writer's laboratory under designedly simple test conditions, and also incidentally to exhibit some of the statistical characteristics of such scores.

EXPERIMENTAL

A 3^2 factorial series of test beverages designated $x \cdot y$ ($x = 0, 1, 2$; $y = 0, 1, 2$) was used in both experiments. The basic material 0.0 was commercial canned tomato juice, while the treatments $x = 1, 2$ and $y = 1, 2$ consisted in the addition of bitter and salt taste stimuli in the amounts of 0.0026 and 0.0052% quinine sulfate and of 0.7 and 1.4% sodium chloride respectively. A volume of each of these nine beverages sufficient for the whole of Experiment A was made up in bulk at the outset. The quantities required for appraisal on different occasions were then stored in sealed glass flasks at 32°F. until used. Test beverages for Experiment B were similarly prepared from a second batch of commercial juice. In both experiments, all test aliquots were submitted for appraisal at 60°F.

Flavor Scoring

For each experiment, 24 untrained subjects were recruited without selection from the scientific, technical and administrative staffs of the main Ottawa laboratories of the National Research Council of Canada. All independently recorded their appraisals of an aliquot of each of the nine test beverages twice, once in a complete set (concurrent appraisal of aliquots of all 9 test beverages) and once in an incomplete set (concurrent appraisal of only 3, 4 or 5 test aliquots). Each subject quantified his or her appraisals of degrees of bitterness by scoring them on a previously described (8) 6-point scale. On this, the integers from 1 to 4 were assigned to successive gradations between "none" (scored 0) and "gross" (scored 5), this last denoting an intensity of psycho-sensory reaction at which the identifying characteristics of bitterness were no longer recognizable. Subjective impressions of saltiness were similarly quantified on an 11-point scale, also previously described (8). On this, a score of 0 was assigned to an intensity considered optimal, scores of -5 and $+5$ respectively denoted extremes of attenuation and intensification at which the identifying characterists of saltiness were unrecognizable, and integer scores of -1 to -4 and $+1$ to $+4$ were again given to intermediate appraisals.

Tasting Schedules

In Experiment A, complete sets C of 9 test aliquots were partitioned

into separately tasted incomplete sets $I.1$, $I.2$, $I.3$, etc. of 3 aliquots each in four different ways:

$I.1$: 0.0, 1.0, 2.0	$I.2$: 0.1, 1.1, 2.1	$I.3$: 0.2, 1.2, 2.2
$I.4$: 0.0, 0.1, 0.2	$I.5$: 1.0, 1.1, 1.2	$I.6$: 2.0, 2.1, 2.2
$I.7$: 0.0, 1.1, 2.2	$I.8$: 0.1, 1.2, 2.0	$I.9$: 0.2, 1.0, 2.1

These partitions, in addition to making possible a variety of incomplete block and treatment interactions, also resulted in the average effects of added sodium chloride, the average effects of added quinine sulfate, and the I and J components of interaction of these two factors (3, sec. 6.15) on the recorded scores being confounded with any differences between tasting sessions in the first, second, third and fourth partitions respectively, and unconfounded in the other three.

Four sessions of tasting of complete and incomplete sets of test aliquots were required of each subject. These took place during mid-morning and mid-afternoon of successive days. Three pairs of subjects each appraised the incomplete sets $I.1$, $I.2$, and $I.3$ in one of three different orders, specified by the rows of a randomly selected 3×3 Latin square. Each of these incomplete sets was consequently tasted before, between and after the others by two subjects. The complete set C was also positioned in the schedule so that it likewise was appraised before, between and after the incomplete sets by two individuals. The actual sequences thus arrived at were:

Subject No.	Tasting session			
	First	Second	Third	Fourth
1, 2	C	$I.2$	$I.3$	I
3, 4	$I.1$	C	$I.2$	$I.3$
5, 6	$I.3$	$I.1$	$I.2$	C

Similarly balanced but independently derived schedules were also specified for the 3 groups of 6 subjects to whom the partitions $I.4 - I.6$, $I.7 - I.9$ and $I.10 - I.12$ were assigned.

In Experiment B, random partitions of the test aliquots into two incomplete sets containing 4 and 5 test aliquots respectively were made. These were tasted in the sequences specified below, where C again denotes a complete set of 9 aliquots, and X an incomplete set of 4, independently randomly selected for each of the 24 subjects.

Subject No.	Tasting session		
	First	Second	Third
25, 26, 31, 32, 37, 38, 43, 44	C	X	$(C - X)$
27, 28, 33, 34, 39, 40, 45, 46	$(C - X)$	C	X
29, 30, 35, 36, 41, 42, 47, 48	X	$(C - X)$	C

In both experiments, the order in which the aliquots constituting any complete or incomplete set were placed was independently randomized for each subject. The subjects were, at the time, unaware that they were participating in a methodological experiment.

Results

The scores for bitterness and saltiness thus recorded for each test aliquot in the complete and incomplete sets of Experiments A and B are listed individually in Tables I and II. Their overall averages for both experiments, which are plotted in Fig. 1 in relation to the composition of the 9 test beverages, clearly increased nearly linearly with

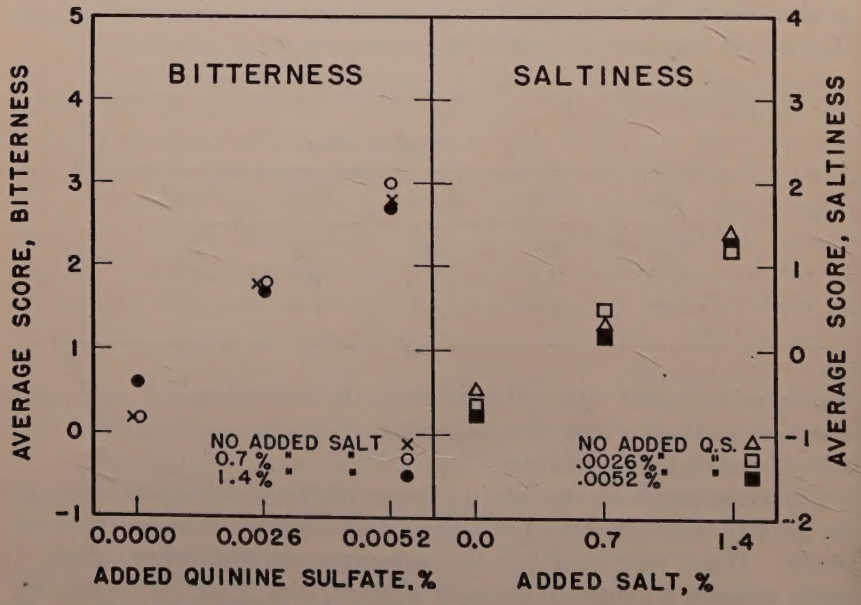


FIG. 1 AVERAGE FLAVOR SCORES IN RELATION TO COMPOSITION OF TEST SAMPLES.

TABLE I

SCORES FOR BITTERNESS AND SALTINESS RECORDED FOR TEST SAMPLES IN COMPLETE (*C*) AND INCOMPLETE (*I*) SETS OF EXPERIMENT A

Subject No.	Score for bitterness																		Score for saltiness															
	0.0		0.1		0.2		1.0		1.1		1.2		2.0		2.1		2.2		0.0		0.1		0.2		1.0		1.1		1.2		2.0		2.1		2.2	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
1	3	0	1	0	1	0	2	2	1	2	0	0	2	2	2	3	1	2	−3	0	2	0	2	2	−2	−1	−2	2	0	1	−3	−2	1	−2	3	−1
2	0	0	0	0	1	0	1	3	3	4	2	1	5	3	4	4	5	3	−1	1	1	2	2	1	−3	0	−2	1	−1	1	−1	−2	1	−2	0	1
3	0	0	0	0	2	0	2	0	0	4	4	3	3	4	3	3	4	5	0	0	2	0	3	2	−2	−3	−1	0	2	−2	2	0	3	1	4	3
4	2	0	0	0	0	0	1	3	3	1	0	0	3	2	4	2	0	2	2	0	−1	1	2	0	0	0	2	0	1	1	−2	0	−3	1	3	2
5	1	0	0	0	0	0	5	0	5	5	2	5	5	4	5	5	5	5	1	1	1	1	5	4	0	3	0	0	3	5	−1	0	0	0	0	5
6	0	0	0	2	0	4	0	2	0	1	1	1	3	2	4	2	2	0	−1	0	0	2	2	4	−2	−1	0	1	3	5	−1	0	0	1	3	3
7	2	0	2	1	3	2	1	0	1	3	5	4	3	3	1	1	2	4	−1	−1	0	2	5	3	−1	−1	3	0	0	2	−1	−1	2	4	5	5
8	0	0	0	0	0	0	5	3	5	4	3	4	4	5	5	4	3	4	0	−1	1	2	3	3	5	1	0	2	0	3	2	3	4	2	2	4
9	0	0	0	0	0	0	0	0	1	1	1	3	5	4	3	4	5	4	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0
10	0	1	0	2	2	0	3	4	3	5	5	0	5	3	3	5	5	3	1	0	0	2	4	3	2	4	2	−2	−1	2	3	3	2	1	3	1
11	0	0	0	0	0	0	2	4	1	4	1	5	4	5	4	5	4	4	0	−2	−2	−2	0	0	−1	0	0	2	0	3	−3	0	0	0	1	0
12	0	0	0	0	0	0	0	2	0	2	0	0	1	2	3	2	2	1	1	0	1	2	3	3	0	1	2	4	3	5	−1	1	1	2	3	3
13	0	1	0	0	2	0	1	1	1	0	0	1	1	2	1	4	0	4	−1	0	1	0	0	0	−1	−1	−2	0	1	1	0	−1	0	2	0	2
14	0	0	0	0	3	1	4	4	3	3	3	5	4	4	5	5	5	4	−1	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	−1
15	0	0	0	0	0	0	0	0	1	0	0	0	3	2	3	2	0	2	0	1	1	1	2	3	0	−1	2	2	4	3	0	0	0	2	2	4
16	0	0	1	1	3	2	0	0	1	1	2	1	0	0	2	1	2	2	0	0	0	1	1	1	−1	0	1	1	2	2	−2	−1	2	1	2	3
17	0	0	0	0	0	0	2	2	1	1	1	1	1	3	3	3	1	3	0	0	0	0	1	1	−1	0	0	0	0	0	−1	−1	1	−1	1	1
18	0	2	0	0	0	0	0	0	2	0	0	2	3	3	3	4	2	2	−2	0	0	0	1	3	−1	0	−1	1	2	2	−2	0	−1	0	−1	2
19	0	1	0	0	0	1	0	0	0	2	0	0	0	1	0	0	0	2	−2	−2	2	1	−2	3	2	0	2	2	2	3	−3	0	1	2	0	2
20	0	0	0	0	0	0	2	2	2	2	1	2	3	4	3	3	3	4	0	0	0	0	1	1	0	−1	2	0	1	0	0	0	0	0	3	0
21	0	1	0	1	0	0	2	3	1	2	1	1	2	3	2	2	2	4	−1	0	0	1	1	1	0	−2	0	0	0	1	−1	0	−1	2	0	0
22	0	0	0	0	0	0	0	1	1	0	2	1	2	2	2	2	1	4	0	0	1	−1	3	4	−2	−2	2	3	3	4	0	2	0	−1	4	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	−1	1	1	0	0	0	0	1	−1	−1	0	2	0	0	3
24	0	1	0	0	0	0	2	3	2	3	1	4	4	4	3	5	3	4	0	0	0	1	1	1	−3	−3	−2	−2	0	5	−3	−5	−1	−3	−1	−3
Average	0.3	0.3	0.2	0.3	0.7	0.4	1.5	1.6	1.6	2.1	1.5	1.8	2.8	2.8	2.8	3.0	2.4	3.0	0.3	0.1	0.5	0.7	1.7	2.0	0.5	0.3	0.3	0.8	1.1	1.9	8.0	0.2	0.6	0.5	1.5	1.6

TABLE II

SCORES FOR BITTERNESS AND SALTINESS RECORDED FOR TEST SAMPLES IN COMPLETE (*C*) AND INCOMPLETE (*I*) SETS OF EXPERIMENT B

Subject No.	Score for bitterness																Score for saltiness															
	0.0		0.1		0.2		1.0		1.1		1.2		2.0		2.1		2.2		0.0		0.1		0.2		1.0		1.1		1.2		2.0		2.1		2.2	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
25	0	0	0	0	0	0	3	1	0	0	4	2	1	3	5	0	5	0	−2	−1	0	0	3	1	−3	0	0	3	0	0	−2	0	−3	0	2	1
26	0	0	0	0	0	0	2	0	0	2	0	0	4	2	2	2	3	3	−3	0	−1	0	1	0	−3	−1	0	2	2	−1	−2	−2	−2	−2	3	3
27	0	0	0	0	0	0	4	3	4	3	4	3	5	5	5	5	5	5	0	−1	0	0	1	1	0	0	0	0	1	1	0	−2	0	0	1	1
28	0	0	0	0	1	0	2	1	2	1	1	1	3	4	5	5	5	4	0	0	0	0	2	1	−2	0	0	0	3	2	0	−1	0	0	3	2
29	1	0	2	1	1	0	2	3	2	3	3	3	2	2	3	2	2	0	0	1	1	2	−3	3	−2	−3	2	1	3	−2	−3	0	−2	1	2	3
30	0	0	1	0	0	1	0	0	1	0	1	1	2	1	2	1	2	2	−1	0	0	2	0	2	−1	0	0	2	1	1	−1	−1	0	1	1	2
31	0	0	1	0	0	0	4	2	2	2	1	3	4	1	5	3	3	1	−1	0	−2	1	0	0	2	−2	2	3	1	3	−2	1	2	−1	3	1
32	0	0	1	0	2	2	2	2	2	1	3	1	1	3	3	2	2	3	0	0	−1	−1	−2	2	−1	−1	2	0	1	0	3	0	0	1	−1	3
33	1	0	0	0	0	0	2	0	3	0	0	3	2	1	3	1	1	1	−1	0	0	0	0	0	2	0	−3	1	4	3	−1	2	−1	1	1	2
34	0	0	0	0	1	0	0	1	1	1	0	0	3	0	5	3	0	1	0	0	0	0	0	0	2	0	0	0	0	1	2	0	0	1	0	0	0
35	0	0	0	0	1	3	1	4	1	1	4	1	3	3	2	4	3	3	−3	−2	0	0	3	3	−1	−1	3	−1	1	4	−3	0	−2	1	3	0
36	1	1	0	0	0	1	4	5	4	5	4	0	4	0	2	5	5	0	−1	−1	0	0	1	1	−3	−5	−4	−5	−4	−3	−4	−5	−2	−5	−5	−4
37	0	0	0	0	3	4	0	1	1	1	4	0	3	0	1	2	2	3	−1	0	0	0	0	1	−2	−3	0	0	0	0	−2	−2	0	1	0	2
38	0	0	0	0	1	0	2	1	3	2	1	2	3	4	4	3	3	1	0	0	2	1	3	2	0	0	1	2	2	1	1	−2	2	0	4	3
39	1	0	0	0	0	0	1	1	2	2	1	1	2	2	2	3	2	4	0	0	0	0	1	1	0	−1	0	0	2	0	−1	−1	−1	1	1	0
40	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	2	1	1	−2	−1	0	0	0	1	−1	0	0	0	1	1	−2	0	0	0	1	1
41	0	0	0	0	0	0	3	3	3	3	2	2	4	4	4	4	4	4	0	0	0	0	1	0	1	−2	0	0	0	1	−2	−1	−2	−2	0	−1
42	0	0	1	0	0	1	1	3	1	2	2	2	3	3	2	5	1	2	0	0	0	−1	0	−1	1	−1	1	−2	−1	−2	−1	−1	1	−2	−1	−3
43	0	0	0	1	0	0	2	2	2	0	1	2	3	3	3	3	1	2	−1	−2	0	0	1	0	−1	0	1	−1	1	2	−2	1	0	1	2	2
44	0	0	2	0	0	2	2	4	2	1	5	5	5	5	4	5	5	4	0	0	−1	0	0	1	−3	0	1	1	0	1	1	0	−3	0	−2	1
45	0	0	0	0	0	1	1	2	1	2	2	3	3	4	4	4	2	3	−1	−2	0	0	1	2	−1	−1	1	0	0	1	−1	−2	−1	1	1	1
46	0	0	0	0	0	1	3	5	1	3	1	1	3	5	2	4	4	4	−1	−4	−1	−1	3	1	−2	−3	0	2	5	0	−5	−2	−2	1	4	5
47	0	0	0	0	0	0	1	4	1	2	1	1	2	3	3	4	2	5	0	0	0	0	1	1	−1	0	0	0	0	0	0	−1	−1	0	0	1
48	0	0	0	1	0	0	3	4	3	3	1	2	3	3	3	3	4	4	−1	−1	0	2	1	1	0	5	2	3	3	0	0	1	2	1	3	2
Average	0.2	0.0	0.3	0.1	0.4	0.8	1.9	2.2	1.8	1.7	2.0	1.6	2.9	2.6	3.1	3.1	2.8	2.5	−0.8	−0.6	−0.1	−0.2	0.8	1.1	−0.9	−0.8	−0.4	−0.5	1.1	0.6	−1.2	−0.8	−0.6	−0.0	1.1	1.2

TABLE III

SCORE TOTALS AND CONTRASTS FOR BITTERNESS AND SALTINESS RECORDED FOR COMPLETE (*C*) AND INCOMPLETE (*I*) SETS OF SAMPLES IN EXPERIMENT A

Subject No.	Score total and contrasts for bitterness																	
	T		*BL*		*BQ*		*SL*		*SQ*		*BL* × *SL*		*BL* × *SQ*		*BQ* × *SL*		*BQ* × *SQ*	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
1	13	11	0	7	4	−1	−5	−2	1	−4	1	0	−3	−2	1	4	1	2
2	21	18	13	10	3	−6	2	−2	0	−6	−1	0	1	−2	−1	4	9	6
3	18	19	8	12	0	−2	5	4	9	−2	−1	1	−1	3	−1	−5	−9	13
4	13	10	5	6	1	−2	−6	−3	−8	1	−1	0	−7	0	−3	6	7	−2
5	28	24	14	14	−8	−6	−4	6	−2	−6	1	1	−1	−1	5	−9	7	9
6	10	14	9	−2	7	2	0	1	−2	−1	−1	−6	−3	−2	−3	4	−5	−4
7	20	18	−1	5	−1	−3	4	7	8	3	−2	−1	2	5	−8	−5	−4	9
8	25	24	12	13	14	−9	−3	0	−5	0	−1	−1	−3	1	3	−3	1	3
9	15	16	13	12	9	4	1	3	3	1	0	0	4	0	−2	−6	6	−2
10	26	23	11	8	−7	−4	4	−5	8	−13	−2	1	2	−1	−2	7	2	5
11	16	27	12	14	4	−12	−1	0	1	0	0	−1	0	−1	2	−3	−2	−3
12	6	10	6	6	6	−2	1	−4	−3	−2	1	−2	−3	0	1	2	−3	4
13	6	13	0	9	0	7	0	1	0	1	−3	3	−3	−3	3	1	3	−5
14	27	26	11	12	−3	−10	3	2	3	2	−2	−1	−4	−3	6	−1	0	−7
15	7	6	6	6	4	6	−3	0	−5	0	−3	0	−3	0	−3	0	1	0
16	11	8	0	0	2	2	7	5	−1	−1	−1	0	−3	0	1	2	−1	2
17	9	13	5	9	−3	1	−1	−1	−3	1	0	0	−4	0	2	2	−6	−2
18	10	13	8	7	4	7	−1	−1	−5	1	−1	1	−1	−5	−1	−7	7	−5
19	0	7	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	13
20	14	17	9	11	−1	−1	−1	0	−1	2	0	0	0	2	2	0	2	2
21	10	17	6	7	−2	−1	−1	−2	1	2	0	2	0	4	2	4	−2	2
22	8	10	5	8	−1	4	1	2	−1	4	−1	2	−1	2	−5	2	−1	−2
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	15	24	10	12	0	−6	−2	0	0	0	−1	1	1	−3	1	−3	3	−3
Average	13.7	15.3	6.8	7.8	0.2	−1.3	0.0	0.5	−0.1	−0.7	−0.8	0.0	−1.3	−0.2	0.0	−0.1	0.7	1.5

TABLE III (*continued*)

Subject No.	Score total and contrasts for saltiness																	
	T		*BL*		*BQ*		*SL*		*SQ*		*BL* × *SL*		*BL* × *SQ*		*BQ* × *SL*		*BQ* × *SQ*	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
1	−2	−1	0	−7	10	−7	13	5	−5	−1	1	−1	3	−1	7	−1	−11	11
2	−4	3	−2	−7	14	−3	6	4	−4	0	−2	3	−2	5	0	1	−4	3
3	13	1	4	2	16	16	9	6	1	−2	−1	1	1	−1	−3	3	−5	13
4	4	5	−5	2	−5	2	6	3	10	−1	5	2	1	2	3	0	19	−4
5	9	19	−8	−1	0	−5	8	10	6	16	−3	2	−5	2	−1	4	−3	−8
6	4	15	1	−2	1	0	12	13	4	3	1	−1	1	1	−3	−5	1	−3
7	12	13	2	4	6	10	13	13	−3	−5	0	2	−4	−2	10	4	18	−8
8	17	19	4	5	2	1	−2	7	2	1	−3	−3	−5	5	13	1	−13	1
9	0	4	0	−4	0	4	0	3	0	1	0	−3	0	−1	0	3	0	1
10	16	14	3	0	7	2	0	−1	4	11	−3	−5	−3	3	9	5	13	−19
11	−4	1	0	4	−4	−14	6	5	2	1	4	−2	−6	−2	0	−4	2	4
12	13	21	−2	1	−2	−9	9	9	1	−3	2	−1	−2	1	0	−3	4	3
13	−2	3	0	3	4	3	3	5	1	−3	−1	3	3	−3	−3	−1	−11	−3
14	1	1	−1	−1	1	−5	1	−1	−5	−5	−1	−1	5	−1	1	−1	−5	7
15	11	15	−1	1	−7	3	8	10	2	0	0	2	2	−2	−4	−2	2	6
16	5	8	1	1	−1	−1	8	7	−4	−1	3	3	−5	1	−1	1	−1	−1
17	1	0	0	−2	4	0	4	3	−2	3	1	1	−3	1	1	3	1	3
18	−5	8	−3	−1	−5	−1	7	7	1	5	−2	−1	0	−1	−2	1	−8	5
19	2	11	0	2	−16	−4	3	10	−13	−4	3	−3	3	−1	3	1	−13	−1
20	7	0	2	−1	−2	3	5	2	1	0	2	−1	2	−1	2	−1	10	3
21	−2	3	−2	0	−2	6	3	4	1	−6	−1	−1	1	−3	3	−5	1	−3
22	11	9	0	−2	2	−6	12	8	2	6	1	−6	3	−2	−3	−10	11	18
23	6	2	−3	3	3	5	1	3	−3	5	2	2	−6	0	−2	6	−6	8
24	−9	−9	−6	−13	6	−9	6	11	0	3	1	1	−3	−1	−3	−13	−3	−15
Average	4.3	6.9	−0.7	−0.5	1.3	−0.4	5.9	6.1	0.0	1.0	0.4	−0.3	−0.8	0.0	1.1	−0.5	0.0	0.9

TABLE IV

SCORE TOTALS AND CONTRASTS FOR BITTERNESS AND SALTINESS RECORDED FOR COMPLETE (*C*) AND INCOMPLETE (*I*) SETS OF SAMPLES IN EXPERIMENT B

Subject No.	Score Total and Contrasts for Bitterness																	
	T		*BL*		*BQ*		*SL*		*SQ*		*BL* × *SL*		*BL* × *SQ*		*BQ* × *SL*		*BQ* × *SQ*	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
13	18	6	11	3	−3	−3	5	−2	3	6	4	−3	−4	3	2	−5	−18	−3
14	12	9	9	7	3	3	−3	1	3	−3	−1	1	3	1	3	1	3	9
15	27	24	15	15	−9	−3	0	0	0	0	0	0	0	0	0	0	0	0
16	19	16	12	13	4	7	2	0	−2	−2	1	0	−3	−2	5	0	1	−2
17	18	16	3	1	−3	−11	1	0	−3	−2	0	−4	0	−2	−2	0	−6	−2
18	9	6	5	3	3	3	1	3	−3	3	0	0	2	0	−2	0	0	0
19	20	12	11	5	−1	−9	−4	1	−4	−3	−1	0	−1	−4	5	−2	−7	−6
20	16	14	3	6	−5	2	4	1	−2	5	−1	−2	−3	0	1	4	−5	2
21	12	6	5	3	−3	−3	−4	3	−6	3	0	0	−4	0	2	−6	6	−6
22	10	6	7	4	7	0	−2	0	−8	−6	−4	1	−8	−5	−2	3	−2	−3
23	15	19	7	7	−3	1	4	0	6	4	−1	−3	1	−5	−5	9	−3	−5
24	24	17	10	3	−12	−13	0	−5	6	+13	2	0	4	−12	0	10	6	2
25	14	11	3	1	−1	5	6	6	8	2	−4	−1	0	−5	−6	9	2	5
26	17	13	9	8	−1	−2	0	−2	−4	−2	−1	−3	−3	−1	3	−5	5	1
27	11	13	5	9	−1	1	−1	2	−1	−2	1	2	−1	0	−1	2	5	4
28	4	4	3	4	1	4	1	0	1	−2	0	0	0	−2	−2	0	−2	−2
29	20	20	12	12	−4	−4	−1	−1	−1	−1	0	0	0	0	2	2	2	2
30	11	18	5	9	−1	−3	−1	−1	−1	−3	−2	−2	2	−6	−4	2	−4	−6
31	12	13	7	7	−3	1	−3	−1	−3	1	−2	−1	−2	1	0	−1	0	−11
32	25	26	12	12	−2	−4	3	2	1	8	0	−3	6	−3	−6	−1	−8	−13
33	13	19	9	10	1	−2	0	1	−2	1	−1	−2	−3	−2	−3	−2	−5	−2
34	14	23	9	12	−1	−4	−1	−4	5	2	1	−2	3	0	5	8	−1	2
35	10	19	7	12	1	−2	0	−1	−2	1	0	2	−2	0	0	8	−2	−2
36	17	20	10	9	−4	−7	−1	−1	−1	−1	1	1	1	3	5	5	5	−1
Average	15.3	14.6	7.9	7.3	−1.5	−1.8	0.3	0.1	−0.4	−0.2	−0.3	−0.8	−0.5	−1.7	0.0	1.7	−1.2	−1.5

TABLE IV (*continued*)

Subject No.	Score Total and Contrasts for Saltiness																	
	T		*BL*		*BQ*		*SL*		*SQ*		*BL* × *SL*		*BL* × *SQ*		*BQ* × *SL*		*BQ* × *SQ*	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*	*C*	*I*
13	−5	4	−4	1	4	−5	12	3	4	−5	−1	−1	5	1	3	3	13	13
14	−5	−1	2	−1	−2	−1	14	5	4	−1	1	5	5	5	−1	5	7	17
15	3	0	0	−1	0	−3	3	6	3	0	0	1	0	−1	0	3	0	−3
16	6	4	1	0	3	−2	10	6	6	4	1	2	1	0	−5	0	3	−2
17	−2	6	−1	−2	−11	18	7	6	−5	−6	8	1	8	1	−8	3	4	15
18	−1	9	1	−2	−1	0	5	6	−1	−6	1	1	1	1	−1	3	−1	3
19	5	6	6	0	−10	−6	5	5	−1	−3	4	0	−6	6	8	−10	2	12
20	1	4	5	3	−5	7	−4	6	−2	4	−2	1	2	−3	−10	3	10	7
21	1	9	0	5	−8	−3	5	3	13	3	1	0	3	2	−1	−6	−23	0
22	2	4	1	−2	−1	−2	1	4	−1	4	0	−2	−2	−2	−2	−2	−4	−2
23	1	4	−2	0	−8	−2	14	10	−2	4	0	−5	4	−3	8	−5	16	−11
24	−22	−27	−11	−14	11	12	0	5	−4	3	−3	−1	−5	1	3	−1	−7	−3
25	−5	−1	−1	0	1	8	5	8	−5	−4	1	3	−1	−3	−1	−1	1	5
26	15	7	2	−2	6	−2	8	8	0	−2	0	3	2	1	2	5	0	7
27	2	0	−2	−1	−4	3	5	3	5	−3	1	0	1	−4	−1	0	−1	0
28	−3	2	1	1	−3	−1	7	4	−3	2	1	−1	1	1	1	1	−3	−1
29	−2	−5	−5	−4	−5	−2	2	3	4	1	1	0	1	2	5	−6	1	4
30	0	−13	−1	−4	−3	2	−2	−4	−6	2	0	−1	−4	−1	4	−1	0	−1
31	1	3	0	6	−2	0	8	5	−2	3	2	−1	0	3	2	−1	4	−9
32	−7	4	−3	0	−1	−2	0	3	2	1	−3	0	3	0	−9	0	17	4
33	−1	0	−1	0	−1	0	5	9	−1	−3	0	−1	2	−3	2	3	8	−3
34	1	−1	−4	8	−8	2	20	15	10	−7	5	2	−1	2	−1	6	1	14
35	−1	1	−2	−1	2	1	2	3	2	1	−1	1	1	−1	−1	3	5	1
36	10	14	5	2	−5	−10	8	−2	−2	−4	1	−1	−1	5	−1	13	1	−1
Average	−0.3	1.4	−0.5	−0.3	−2.1	0.5	5.8	5.0	0.8	−0.5	0.8	0.3	0.8	0.4	−0.2	0.8	2.3	2.8

the concentration of added taste stimuli over the experimental range. Some effect of 1.4% added salt on appraisals of bitterness is also suggested.

STATISTICAL ANALYSIS

Some Metrical Characteristics of Scores

Equal increments in score correspond to equal increments in appraised flavor intensity. However, because of differences in sensory thresholds, acuities, or preferred intensities, the relation of scores to specific sensory stimuli may vary appreciably from one subject to another. There may also be discrepancies in repeated appraisals of the same stimulus by the same subject. A single score x_{ijk} recorded by the ith of a group of u subjects for the jth of v test materials at the kth of w repeated appraisals under similar test conditions must generally therefore be regarded as the resultant of components such that

$$x_{ijk} = a + x_{.j.} + y_{ij.} + z_{ijk}$$

Here a denotes the group general mean, $x_{.j.}$ the departure from a of the group average for j, $y_{ij.}$ the average departure from $x_{.j.}$ in scores recorded for j by i due to individual sensory or preference characteristics, and z_{ijk} a random discrepancy affecting i's kth scoring of j. By definition, $\sum_i (y_{ij.}) = 0$ for each j, and $\sum_k (z_{ijk}) = 0$ for each ij. Experience indicates that it cannot be assumed that the discrepancies $y_{ij.}$ for different j are either equally variable or statistically independent.

Skewness, kurtosis and discreteness of distribution are prominent features of the series of scores listed in Tables I and II. They are however less pronounced in the totals and linear score contrasts identified with single degrees of freedom (10, ch. 11) in the customary analysis of variance of a 3^2 experiment, numerical values of which are listed by subjects in Tables III and IV. The contrasts corresponding to the average linear, average quadratic and interaction effects of the two taste factors, denoted by BL, BQ, $BL \times SL$ etc., are specified by the numerical coefficients on page 12.

Computation for each of the 72 score totals and contrasts listed in Table III and IV of the cumulant estimates k_2, k_3, and k_4 and the indices of skewness $| g_1 | = | k_3/k_2^{3/2} |$ and of kurtosis $g_2 = k_4/k_2^2$ gave average $| g_1 | = 0.6$ and average $g_2 = 0.8$. This degree of skewness and leptokurtosis, if uniformly characteristic of all contrasts, would be unlikely seriously to invalidate approximate tests of significance based on "normal parent" 5% or 1% points of t or $F(2, 6)$. It must however also be noted that either unequal between-subject variance

Contrast	Coefficient of sample score in contrast								
	0.0	0.1	0.2	1.0	1.1	1.2	2.0	2.1	2.2
BL	−1	−1	−1	0	0	0	1	1	1
BQ	1	1	1	−2	−2	−2	1	1	1
SL	−1	0	1	−1	0	1	−1	0	1
SQ	1	−2	1	1	−2	1	1	−2	1
BL × *SL*	1	0	−1	0	0	0	−1	0	1
BL × *SQ*	−1	2	−1	0	0	0	1	−2	1
BQ × *SL*	−1	0	1	2	0	−2	−1	0	1
BQ × *SQ*	1	−2	1	−2	4	−2	1	−2	1

or non-zero covariance of the $(y_{ij.} + z_{ijk})$ for different j will result in non-independence of the between-subject within-group variations in *BL*, *BQ* etc. In the present instance, covariation of the specified contrasts of the scores for saltiness may have been negligible, the combined within-experiment correlation coefficients for the four pairs of contrasts *BL* and *BQ*, *SL* and *SQ*, *BL* × *SL* and *BL* × *SQ*, and *BQ* × *SL* and *BQ* × *SQ* when averaged irrespective of sign, being only $r = 0.05$. The corresponding average for the bitterness contrasts was however $r = 0.35$.

Test for Effects of Grouping

Conceivably, the grouping and order of presentation of test samples might have biassed the scores recorded for both complete and incomplete sets, either similarly or differentially. To test these two possibilities, the between-subject variance of both the sum $(C + I)$ and the difference $(C - I)$ of the totals T and contrasts *BL*, *BQ* etc. of the scores in each experiment was analyzed (5, sec. 42) as indicated in Tables V, VIa and VIb. (In this and succeeding analyses, the results for Subject No. 36 were excluded, because investigation revealed that the radical discrepancies between his score contrasts and those recorded by the other 23 participants in Experiment B were due to unreported linguistic difficulties). The mean squares listed in these tables have been reduced to a common single-score basis by division by 18 in the case of T and by twice the sum of the squares of the previously specified contrast coefficients in the case of *BL*, *BQ* etc.

None of the 10 variance ratios $A.1/A.3$, $A.2/A.3$ or $B.1/B.2$ for total score T, listed in Table V, attains its "normal parent" upper 5% point of 3.49, 2.85 or 3.49 respectively. Six of the 128 ratios $A.1(a)/A.3$, $A.1(b)/A.3$, $A.2/A.3$ and $B.1/B.2$ of the mean squares for the score

TABLE V

ANALYSES OF BETWEEN-TASTER VARIANCE OF SUM ($C + I$) AND DIFFERENCE ($C - I$) OF TOTAL SCORES FOR SAMPLES TASTED IN COMPLETE AND INCOMPLETE SETS OF EXPERIMENTS A AND B

Variance associated with	D.f.	Mean sq. (bitter)		Mean sq. (salt)	
		$C + I$	$C - I$	$C + I$	$C - I$
A.1. Different partitions into "incomplete" sets	3	20.55	1.90	16.11	1.91
A.2. Order of tasting "complete" and same "incomplete" sets	8	4.97	1.01	11.49	2.25
A.3. Different tasters of same sets in same order	12	13.31	0.78	8.16	1.84
B.1. Order of tasting "complete" set	2	1.32	8.54	0.04	1.32
B.2. Different tasters and "incomplete" sets, same order of tasting "complete" set	20	6.51	2.49	4.51	3.83

contrasts listed in Tables VIa and VIb exceed their "normal parent" upper 5% points of 4.75, 3.88, 2.85 and 3.49, this number agreeing as closely as is possible with the chance expectation of 6.4 appropriate to 128 independent random variates. The individual probability of the most extreme ratio, namely 5.36 for $B.1/B.2$ for the sum $(C + I)$ of the contrast $BQ \times SQ$ of scores for saltiness, may be calculated from the Incomplete Beta-function tables (9) to be approximately 0.013. This exceeds 1/128, and indeed the probability of the most extreme of 128 independent deviates exceeding its 1.3% point purely by chance would be $1 - (0.987)^{128} = 0.82$. Actually the 128 variance ratios in question are not independent, partly because $A.1(a)/A.3$, $A.1(b)/A.3$ and $A.2/A.3$ for any specified contrast have the same denominator, and partly because of the correlations and variance-heterogeneity noted above. However, even if it is supposed that these factors operate to reduce the recorded aggregate of variance ratios to the statistical equivalent of only half as many independent observations, the corresponding probability $1 - (0.987)^{64}$ would still be as large as 0.43. It may be inferred therefore that in the aggregate these analyses of variance are not indicative of significant bias in the recorded scores attributable either to partition of the 9 test samples into incomplete appraisal sets, or to the order of presentation of the complete and incomplete sets.

TABLE VIa

ANALYSES OF BETWEEN-TASTER VARIANCE OF SUM $(C + I)$ AND DIFFERENCE $(C - I)$ OF SCORE CONTRASTS FOR SAMPLES TASTED IN COMPLETE AND INCOMPLETE SETS OF EXPERIMENTS A AND B

Variance associated with	D.f.	Mean square (bitterness)								Mean square (saltiness)							
		BL		*BQ*		*SL*		*SQ*		*BL*		*BQ*		*SL*		*SQ*	
		$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$
A.1. Different partitions into "incomplete" sets																	
(a) between inter and intra "incomplete" set contrasts	1	10.45	0.03	3.06	1.25	1.00	1.83	1.48	0.97	10.89*	0.10	0.05	0.13	10.02	3.76	2.79	0.07
(b) between intra "incomplete" set contrasts	2	2.95	0.84	0.93	1.20	0.57	0.13	0.26	1.97	1.62	0.26	2.72	2.06	0.16	0.39	0.62	0.30
A.2. Order of tasting "complete" and same "incomplete" sets	8	1.95	1.44	1.25	1.01	1.15	0.91	0.42	0.82	3.13	1.52	3.95	2.95	5.56	1.02	2.51	1.09
A.3. Different tasters of same sets in same order	12	9.16	1.25	2.70	0.58	3.07	1.18	0.73	1.09	2.18	1.48	2.80	1.11	4.00	1.02	1.51	0.60
B.1. Order of tasting "complete" set	2	1.35	2.94	1.16	0.67	0.79	0.74	1.07	0.06	2.70	2.24	0.36	2.20	3.88	0.70	2.45	2.09
B.2. Different tasters and "incomplete" sets, same order of tasting "complete" set	20	4.18	1.43	1.21	1.03	1.22	1.65	0.99	0.94	1.36	2.67	1.13	3.52	5.80	3.91	0.85	1.95

*Ratio to mean square A.3 exceeds "normal parent" 5% point

TABLE VIb

ANALYSES OF BETWEEN-TASTER VARIANCE OF SUM ($C + I$) AND DIFFERENCE ($C - I$) OF SCORE CONTRASTS FOR SAMPLES TASTED IN COMPLETE AND INCOMPLETE SETS OF EXPERIMENTS A AND B

Variance associated with	D.f.	Mean square (bitterness)								Mean square (saltiness)							
		$BL \times SL$		$BQ \times SL$		$BL \times SQ$		$BQ \times SQ$		$BL \times SL$		$BQ \times SL$		$BL \times SQ$		$BQ \times SQ$	
		$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$	$C+I$	$C-I$
A.1. Different partitions into incomplete sets																	
(a) between inter and intra "incomplete" set contrasts	2	0.94	1.51	0.38	0.50	2.23*	0.01	1.11	1.04	0.63	2.77	4.89	1.84	0.05	2.00	0.16	2 16
(b) between "incomplete" set contrasts	1	0.05	0.04	0.90	0.22	2.00*	0.22	0.39	0.01	3.76	3.01	3.78	2.17	1.84*	0.78	0.01	3.77
A.2. Order of tasting "complete" and same "incomplete" sets	8	0.28	0.63	0.79	0.94	0 28	0.72*	0.43	0.60	2.09	0.95	2.59	0.84	0.46	0.97	1.72	1.33
A.3. Different tasters of same sets in same order	12	0.54	0.43	1.27	1.76	0.38	0.22	0.60	1.03	1.10	1.13	1.53	1.11	0.29	0.76	1.83	3.42
B.1. Order of tasting "complete" set	2	0.89	0.42	0.54	4.38	0.55	1.07	0.59	0.52	0.10	4.43	0.32	0.14	0.58	0.27	6.91*	1.32
B.2. Different tasters and "incomplete" sets, same order of tasting "complete" set	20	0.65	1.65	1.05	3.00	0.62	1.31	1.07	0.93	1.47	1.68	1.18	5.71	0.62	1.67	1.29	3.16

*Ratio to mean square A.3 or B.2, respectively, exceeds "normal parent" 5% point.

Mean Scores in Relation to Taste Factors

The means m of the sum $(C + I)$ and difference $(C - I)$ of the recorded totals and contrasts of scores for bitterness and saltiness are listed in columns 2, 3, 8 and 9 of Table VIII. Columns 4, 5, 10 and 11 show the unreduced variance v of these quantities, estimated from 23 d.f. between subjects in Experiment A and from 22 d.f. (excluding Subject 36) in Experiment B, i.e. ignoring groupings of test samples into appraisal sets. Columns 6, 7, 12 and 13 give the resulting Student-Fisher t ratios $\sqrt{24}/m\sqrt{v}$ and $\sqrt{23}m/\sqrt{v}$.

As corresponding t from the two experiments are statistically independent, their indications may be pooled by use of the probability integral transformation (5, sec. 21.1). When applied to their respective "normal parent" probabilities (7) this resulted in the following $-2S(\ln P)$ for the sum $(C + I)$ of the specified score contrasts. Values falling outside the central 95% and 99% chance ranges delimited by the $2\frac{1}{2}$% and $97\frac{1}{2}$%, and by the 1/2% and $99\frac{1}{2}$% points of χ^2(4 d.f.), namely 11.14 and 0.48, and 14.86 and 0.21, are respectively distinguished by single and double asterisks.

	BL	*BQ*	*SL*	*SQ*	*BL* × *SL*	*BL* × *SQ*	*BQ* × *SL*	*BQ* × *SQ*
Bitterness	>46**	0.71	5.17	1.54	0.10**	5.80	0.09**	13.82*
Saltiness	2.08	1.23	>46**	5.50	7.76	4.41	6.70	10.22

Clearly, *BL* for appraisals of bitterness and *SL* for appraisals of saltiness are both highly significant ($P < .001$), while *BQ* for bitterness and *SQ* for saltiness are not. It may be concluded therefore that on the average the recorded scores for these two characteristics increased sensibly linearly with the added concentrations of quinine sulfate and sodium chloride. None of the values of $-2S(\ln P)$ for the six contrasts *BL*, *BQ*, *BL* × *SL* etc. of scores for saltiness of samples actually of the same sodium chloride content but differing in bitterness is extreme, but three of those for the corresponding six contrasts *SL*, *SQ*, *BL* × *SL* etc. of scores for bitterness of samples actually of the same bitterness but differing in saltiness fall outside their central 95% range. Although some allowance must be made here for correlation in different score contrasts from the same experiment, it nevertheless seems reasonable to conclude that appraisals of saltiness were not demonstrably affected by the imposed variations in bitterness, but that the previously sug-

gested partial masking of differences in bitterness at the highest experimental level of saltiness was statistically significant.

A similar combination of the transformed normal parent probabilities of the t for the difference $(C - I)$ of the various score contrasts resulted in the following $-2S(\ln P)$.

	T	BL	BQ	SL	SQ	$BL \times SL$	$BL \times SQ$	$BQ \times SL$	$BQ \times SQ$
Bitterness	8.98	2.49	6.41	1.84	2.46	3.02	1.90	0.67	1.87
Saltiness	0.18**	1.80	3.73	5.14	4.86	7.39	6.07	3.49	2.08

None of these 9 quantities for the bitterness score contrasts falls outside its central 95% interval. Only 4 of the corresponding 9 quantities for saltiness lie outside their central 50% interval, but of these, that for the difference in the score totals T falls just below its 99½% point, having $P = 0.996$. The probability of so extreme a value occurring in 1 or more of 9 independent random variates, namely $1 - \{1 - 2(1 - P)\}^9$, would be about 0.07, so this observation is perhaps suggestive, although certainly not decisively indicative, of a generally slightly higher scoring of the saltiness of the samples appraised in incomplete sets, the recorded average difference being 0.2 scale unit. Here again however there is no indication of any consistent discrepancy in the score contrasts recorded for the samples appraised in complete and incomplete sets.

Inter-experiment discrepancies in average score totals and contrasts were tested for significance, without assuming equality of intra-experiment, intra-subject variance, by Welch's procedure (11). The results were suggestive ($P = 0.02$) of a generally slightly lower score for saltiness in Experiment B than in A, corresponding to an average difference of 0.4 scale unit per sample, but of no consistent difference between corresponding score contrasts.

Inter-subject Variance

It is to be noted from Table VII that the scores recorded for both bitterness and saltiness resulted in a larger between-subject variance of the sums $(C + I)$ than of the differences $(C - I)$ of their complete and incomplete set totals T in both experiments. This is indicative of some consistent individual differences in appraisal of the same taste

TABLE VII

SIGNIFICANCE OF DEPARTURES FROM ZERO OF RECORDED MEANS OF SUM ($C + I$) AND DIFFERENCE ($C - I$) OF SCORE TOTALS AND CONTRASTS FOR SAMPLES TASTED IN COMPLETE AND INCOMPLETE SETS OF EXPERIMENTS A AND B

Total or Contrast	Experiment A						Experiment B					
	Mean, m		Variance, v		$t = m/\sqrt{v/24}$		Mean, m		Variance, v		$t = m/\sqrt{v/23}$	
	$C + I$	$C - I$	$C + I$	$C - I$	$C + I$	$C - I$	$C + I$	$C - I$	$C + I$	$C - I$	$C + I$	$C - I$
Bitterness:												
T	29.00	−1.67	204.3	18.1	—***	1.92	29.43	0.48	108.7	54.7	—***	0.44
BL	14.54	−1.04	74.0	14.7	8.28	−1.33	15.26	0.30	47.1	18.8	11.56	0.48
BQ	−1.13	1.46	74.0	29.1	−0.64	1.32	−1.30	0.22	43.3	36.0	−1.74	0.24
SL	0.50	−0.50	25.1	12.3	0.49	−0.70	0.56	−0.04	14.2	18.8	0.72	−0.07
SQ	−0.75	0.58	22.2	38.4	−0.78	0.46	−0.30	−1.09	35.9	30.9	−0.24*	−1.33
$BL \times SL$	−0.71	−0.79	3.7	4.6	−1.81	−1.81	−1.26	0.39	5.4	12.3	−2.61	0.75
$BQ \times SL$	−0.13	0.13	24.2	31.2	−0.13	0.11	1.35	−1.35	24.0	75.0	1.32	−1.05
$BL \times SQ$	−1.46	−1.04	13.8	9.0	−1.92	−1.70	1.96	0.57	14.7	30.9	−2.45*	−0.70
$BQ \times SQ$	2.13	−0.79	41.4	59.4	1.62	−0.50	−3.17	0.22	73.9	64.0	−1.77	−0.18
Saltiness:												
T	11.21	−2.54	186.3	33.5	—	−2.15*	3.30	−1.91	73.9	64.9	—	−1.61
BL	−1.21	−0.13	34.1	15.9	−1.01	−0.28	0.17	−0.35	17.8	31.8	0.20*	−0.42
BQ	0.96	1.71	113.8	64.6	0.14	1.00	−2.70	−2.70	38.2	122.3	−2.09	−1.65
SL	11.96	−0.21	53.6	13.0	8.00***	−0.15	11.09	1.09	67.5	43.4	6.47***	1.12
SQ	0.96	−1.04	66.0	26.0	0.58	−1.04	0.30	1.61	35.8	70.7	0.77	1.30
$BL \times SL$	0.08	0.67	12.2	10.3	0.12	1.01	1.22	0.61	10.7	15.4	1.78	1.05
$BQ \times SL$	0.58	1.67	60.0	27.0	0.39	1.57	0.52	−1.13	26.5	124.8	0.49	−0.69
$BL \times SQ$	−0.83	−0.75	9.5	22.5	−1.32	−0.77	1.48	0.70	14.7	37.0	1.85*	0.77
$BQ \times SQ$	.0.83	−0.92	112.8	187.1	0.12	−0.33	5.65	−0.35	129.4	215.2	2.38	−0.16

*(***)Exceeds "normal parent" 5% (0.1%) point.

stimuli. The likewise larger between-taster variance of the sum $(C + I)$ than of the difference $(C - I)$ of the major contrasts BL of the scores for bitterness and SL of the scores for saltiness in both experiments is further indicative of consistent individual differences in reaction to the imposed variations in taste stimuli, analogous to a block-treatment interaction. Conversely the smaller between-taster variance of the sum $(C + I)$ than of the difference $(C - I)$ of several of the mostly non-significant non-linear and interaction score contrasts indicates some negative intra-individual correlation of corresponding scores recorded on the two occasions of appraisal of the same test material, resulting from variable identification, by the same individual, of the constituent intensities of bitterness and saltiness in each appraised mixture.

No consistent effect of grouping on the between-subject variance of corresponding score totals or contrasts was evident. In Experiment A, this variance was larger in the complete than in the incomplete set results in 10 instances and smaller in 8. In Experiment B, it was larger in 9 instances and smaller in 9. Fig. 2 shows the individual

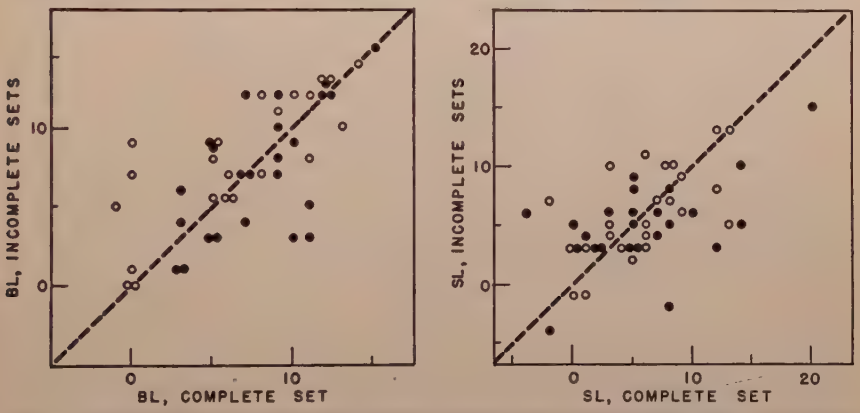


FIG. 2. RELATION OF INDIVIDUAL SCORE CONTRASTS RECORDED FOR COMPLETE AND INCOMPLETE SET APPRAISALS. HOLLOW CIRCLES, EXPERIMENT A; SOLID CIRCLES, EXPERIMENT B.

contrasts BL and BS recorded for incomplete sets in both experiments in relation to their corresponding complete set values. Brought to a single-score basis, the average between-subject variance of score contrasts corresponded to a standard deviation of 1.1 unit for bitterness on the 6-point scale and of 1.3 unit for saltiness on the 11-point scale.

DISCUSSION

The objective of the scoring procedure considered here is limited

to quantification by scoring on an interval (11-point) or ratio (6-point) scale (4), of subjective estimates of the intensity of perceived flavor sensations. This implies measurement of subjective reactions, as distinct from sensory stimuli. However, the current experiments are in accordance with others previously described (8) in indicating a consistent relation of group average scores to known concentrations of such stimuli. In the present instance, moreover, the score contrasts thus obtained were not demonstrably biassed by groupings of samples into incomplete sets for appraisal.

When characteristic individual reactions lead to score differences analogous to "complete block $\times$ treatment" interactions, adoption of experimental designs in which individual subjects correspond to "complete blocks" is evidently desirable. This is compatible with grouping of a series of test samples into incomplete sets for concurrent appraisal, and groupings of a factorial series providing for balanced partial confounding may possibly result in localization as well as in detection of any average differences in scoring between incomplete sets.

The power and simplicity of customary additive-model analysis of variance techniques are lost when a significant block-treatment interaction indicates non-additivity of block and treatment effects. Inferences respecting specified treatment contrasts may as here still be possible from t or F tests interpreted with due caution, but if non-orthogonality of different contrasts is appreciable an overall test of a general null hypothesis of zero means for all contrasts may involve computation of multivariate statistics such as Hotelling's T or Wilks' likelihood ratio criterion (12, chap. 7). This was unnecessary here because of the undoubted significance of BL and SL considered individually.

If test subjects are randomly recruited from some definable population and have no special training or experience, recorded group means and variances which are functions of the group parameters a, $x_{\cdot j\cdot}$, y_{ij} and z_{ijk} may be regarded statistically as unbiassed estimates of identical functions of corresponding population parameters α, $\xi_{\cdot j\cdot}$, η_{ij}, and ζ_{ijk}. By selection and training, enhanced $x_{\cdot j\cdot}$ and reduced $y_{ij\cdot}$ and z_{ijk} might be obtained, but the practical interpretation of mean scores and score contrasts would then require knowledge of the specific relation of the group a and $x_{\cdot j\cdot}$ to α and $\xi_{\cdot j\cdot}$. Moreover, unless a large number of qualitatively similar appraisals must be made, it is possible that the precision of group averages might be increased more economically by employment of additional untrained subjects.

ACKNOWLEDGEMENTS

Collaboration of the 48 subjects who participated in these experiments under the technical supervision of Elinor M. Hamilton is gratefully acknowledged. Miss Hamilton, together with Mrs. Florence H. Suddon also assisted largely in the numerical analyses of the resulting scores.

REFERENCES

1. Boggs, Mildred M. and Hanson, Helen L. Analysis of foods by sensory difference tests. *Advances in Food Research*, 2:219–258. Academic Press, Inc., New York. 1949.
2. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3:22–38. 1947.
3. Cochran, W. G. and Cox, Gertrude M. *Experimental Designs*. John Wiley and Sons, Inc., New York. 1950.
4. Coombs, C. H. Mathematical models in psychological scaling. *J. Amer. Statist. Assoc.* 46:480–489. 1951.
5. Fisher, R. A. *Statistical Methods for Research Workers*. 10th ed. Oliver and Boyd. Edinburgh. 1946.
6. Gayen, A. K. The distribution of the variance ratio in samples of any size drawn from non-normal universes. *Biometrika* 37:236–255. 1950.
7. Hartley, H. O. and Pearson, E. S. Table of the probability integral of the *t*-distribution. *Biometrika*, 37:168–172. 1950.
8. Hopkins, J. W. A procedure for quantifying subjective appraisals of odor, flavor and texture of foodstuffs. *Biometrics*, 6:1–16. 1950.
9. Pearson, K. (ed.). *Tables of the Incomplete Beta-function*. Biometrika Office, University College, London. 1934.
10. Snedecor, G. W. *Statistical Methods*. 4th ed. The Collegiate Press, Inc. Ames, Iowa. 1946.
11. Welch, B. L. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34:28–35. 1947.
12. Rao, C. R. *Advanced Statistical Methods in Biometric Research*. John Wiley and Sons, Inc., New York. 1952.

SOME STATISTICAL METHODS IN TASTE TESTING AND QUALITY EVALUATION[a,b]

RALPH ALLAN BRADLEY

The Virginia Agricultural Experiment Station
Virginia Polytechnic Institute
Blacksburg, Virginia

1. INTRODUCTION

1.1 *Introductory Remarks.*

The title suggested for this paper is a general one and the discussion which follows is necessarily of a rather broad nature. There is a definite need for improved communication between the consumer of statistical methods, in this case the food technologist, the home economist and the horticulturist, on the one hand and the manufacturer of statistical methods, the mathematical statistician, on the other. The difficulty is accentuated by the general lack of mathematical training for the agricultural sciences and the sometimes aloofness of the statistician.

The mathematical statistician has tended to publish his research in concise mathematical style for an audience consisting principally of mathematical statisticians. Eventually the statistical methods, if they are good ones, become translated for use in applied problems. The time lag in some cases is considerable and should be decreased or eliminated. This can be accomplished in several ways. The ideal method would seem to involve the publication of an applied paper as a companion to the usual paper setting forth the theory of a new method. Lacking this, it is at least necessary that the research worker have some means of understanding of or reference to new statistical procedures. It is to this purpose that this paper may be of some small value.

---

[a]Prepared under a Research and Marketing Act contract, Project RM:c-629.1, with the Bureau of Agricultural Economics.

[b]This paper was presented to the Joint Symposium of The Biometric Society and the American Society of Horticultural Science at Cornell University, September 9, 1952.

1.2 *Bibliography.*

A fairly extensive classified bibliography is appended to this paper. No claim is made that it is exhaustive, which it obviously is not, nor that it affords a complete classification. The references given have been selected with a view to showing typical illustrative examples of statistical methods used in taste testing and procedures which are thought to be applicable in taste testing and quality evaluation.

Considerable assistance was derived from two available lists of references, [62] and [67]. The bibliography here is a condensation of abstracted references which were reported to the Bureau of Agricultural Economics [63, 64, 65, 66] and which received limited circulation. Many references are indicated which do not receive consideration in the body of this paper.

It is hoped that the bibliography provided may assist in bridging the gap between statistician and food technologist.

1.3 *Types of Taste Panels and Their Purposes.*

Taste panels may be classified in four types and exist for reasons which are primarily different.

(i) *Taste Panels for the Detection of Differences.*

These panels are usually small and it seems preferable to have from three to ten good judges to a larger untried panel. This sort of panel is one which is used for research purposes only. Rather intensive training of panel members is usually undertaken but in some cases it is not necessary that the members of a panel agree on their preferences or even on their judgments. It is necessary that a judge demonstrate ability to repeat his judgments.

For the development of new products, the improvement of old ones and the detection of insecticides and adulterants or effects of packaging or storage one is concerned only with the question of the existence of true differences. Unfortunately, presence or absence of differences is confounded with the presence or absence of taste acuity on the part of the judges.

(ii) *Taste Panels for Quality Control.*

Taste panels for quality control are usually panels of long standing and of more experience than the first type. Such panels are usually used for the maintenance of standards and as such are interested in the lack of differences or variability of some few specified products. Chart records may be kept both of the day-by-day performance of

individual judges and of the panel comparisons as a whole. Again, these panels do not do preference testing in any way. Taste panels for quality control may be quite small but must be efficient.

(iii) *Taste Panels for Consumer Preference.*

In consumer preference testing panels are large and untrained. Usually no standards are provided and decisions are based on preferences alone. To be useful such panels must be representative of the consumer market of interest. Test procedures should be kept simple and the number of items compared should be small.

(iv) *Taste Panels for Quality Evaluation.*

Taste testing for quality evaluation is usually one phase of a more elaborate evaluation procedure. Composite quality scores consist of weighted averages of a variety of determinations. This kind of taste procedure is used in certain United States Standards for Grades. The tasting is usually done by a very small number of official graders. An attempt is made to conform to a uniform scoring system over long periods of time. Interest is in an absolute taste score and not in comparative scores for several products as is usually sufficient in the other types of panel testing.

Problems of considerable interest arise from the proper weighting of attribute measurements in quality evaluation.

Fundamental to any work with taste panels are problems in the selection of the panel and the choice of experimental designs and scoring techniques. We turn to a consideration of these problems.

II. THE SELECTION OF A TASTE PANEL

2.1 *General Discussion.*

Triangle tests, wherein a taster is required to pick the odd sample from a set of three samples, have been widely used for the selection of a taste panel. The procedure has the advantage of simplicity. In some cases each potential panel member is required to perform a predetermined number of these comparisons and the best judges of the group are selected for a panel. Concern has existed regarding the quality of the judges so obtained and the number of triangle tests which should be given. In general insufficient testing is done due to the time consumed and limitations on suitable experimental material.

The author has recently learned of attempts by food technologists to develop sequential testing and selection procedures and that without regard to the systems of sequential analysis developed by Wald [85] and Rao [84]. In applying a sequential analysis control is obtained

over the quality of judges selected but simple applications do not necessarily lead to the selection of the best judges available. It is thought that sequential methods will provide considerable improvements over most selection procedures now used and constitute a saving of time and material.

Lombardi [83] in a thesis developed the applications of both the Wald and Rao methods of sequential analysis to the use of triangle tests in judge selection.

2.2 *Wald's Sequential Analysis Applied.*

Let p be the true proportion of correct decisions in triangle tests if the judge could continue testing indefinitely. This may be thought of as the judge's inherent ability under the test administered. Judges having abilities less than p_0 will be ruled unacceptable for a panel and those with abilities greater than p_1 will be selected.

A graph is drawn on which the number of observations m is plotted as abscissa and the number of correct decisions d_m is specified as the ordinate. Two lines L_0 and L_1 , having a common slope, divide the graph into acceptance and rejection regions. The slope and intercepts of L_0 and L_1 are given by Wald [85] (c.f. section 5.3.3) and they depend on the specification of p_0 and p_1 and parameters α and β. α is defined as the probability of selecting an unacceptable judge and β is the probability of rejecting an acceptable judge. p_0 , p_1 , α, β are at the disposal of the experimenter. If potential judges are in good supply, he may take α small and β large.

The common slope of L_0 and L_1 is given by

$$(1) \qquad s = \frac{\log \dfrac{1 - p_0}{1 - p_1}}{\log \dfrac{p_1}{p_0} - \log \dfrac{1 - p_1}{1 - p_0}}$$

and intercepts h_0 and h_1 respectively are

$$(2) \qquad h_0 = \frac{\log \dfrac{\beta}{1 - \alpha}}{\log \dfrac{p_1}{p_0} - \log \dfrac{1 - p_1}{1 - p_0}}$$

and

$$(3) \qquad h_1 = \frac{\log \dfrac{1 - \beta}{\alpha}}{\log \dfrac{p_1}{p_0} - \log \dfrac{1 - p_1}{1 - p_0}}$$

Before a definite decision is reached on the specification of p_0, p_1, α, β it is useful to compute the average number of tests which will be required. This depends on the ability of the judge p and the test specification. A rough plot or table can be computed by considering special values of p as shown by Wald (c.f. section 5.5). $E_p(n)$ is the average sample number for a judge of ability p. Then,

(4) for $p = 0$ (no ability)

$$E_0(n) = \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{1-p_1}{1-p_0}},$$

(5) for $p = p_0$ (maximum unacceptable ability)

$$E_{p_0}(n) = \frac{(1-\alpha)\log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha}}{p_0 \log \frac{p_1}{p_0} + (1-p_0)\log \frac{1-p_1}{1-p_0}},$$

(6) for $p = p_1$ (minimum acceptable ability)

$$E_{p_1}(n) = \frac{\beta \log \frac{\beta}{1-\alpha} + (1-\beta)\log \frac{1-\beta}{\alpha}}{p_1 \log \frac{p_1}{p_0} + (1-p_1)\log \frac{1-p_1}{1-p_0}},$$

and

(7) for $p = 1$ (infallible ability)

$$E_1(n) = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_0}}.$$

One further average sample number may sometimes be computed and this case occurs when $p = s$, the slope of the control lines. Then,

(8) $$E_s(n) = -\frac{\left(\log \frac{\beta}{1-\alpha}\right)\left(\log \frac{1-\beta}{\alpha}\right)}{\log \frac{p_1}{p_0} \log \frac{1-p_0}{1-p_1}}.$$

An example has been constructed in which judges with inherent abilities .33 and .60 have been simulated. The test specification used the values $p_0 = .40$, $p_1 = .65$, $\alpha = .05$ and $\beta = .05$. These values are not necessarily suitable for all judge selection and in general p_0 and p_1

are too small while β may be sometimes profitably increased. Average sample numbers are shown in Table I.

TABLE I
AVERAGE SAMPLE NUMBERS FOR THE TRIANGLE TEST

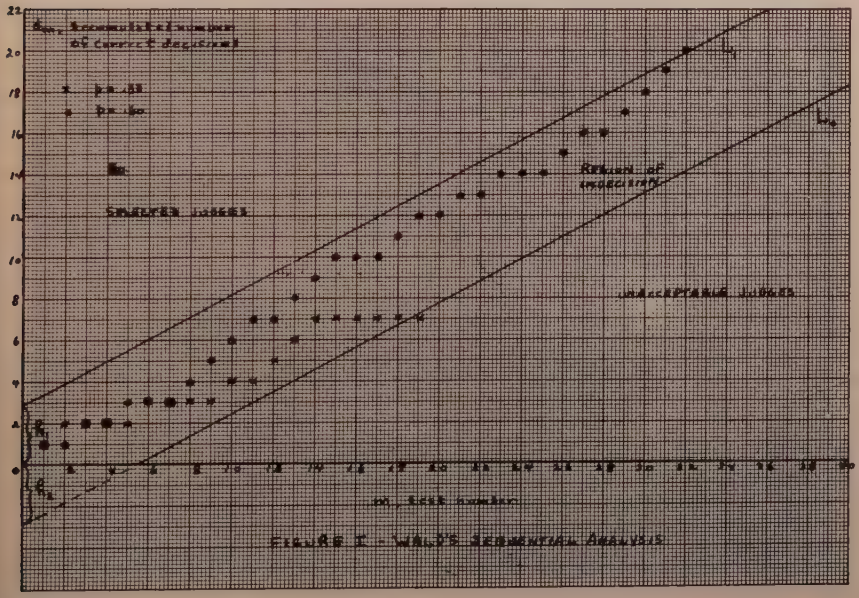
p	$E_p(n)$
0	5
0.40	21
0.65	21
1	6

Substitution in (1), (2) and (3) yields $s = .53$, $h_0 = -2.87$ and $h_1 = 2.87$. The equations of L_0 and L_1 become

$$L_0 : \quad d_m = .53m - 2.87 \quad \text{and} \tag{9}$$

$$L_1 : \quad d_m = .53m + 2.87.$$

The diagramatic procedure is shown in Figure I for two judges of



abilities .33 and .60. The analysis rejected the first of these judges but accepted the second. In so doing a judge has been obtained with

ability somewhat below that desired but one who was better than those which were ruled to be unacceptable.

2.3 *Rao's Sequential Analysis.*

Rao [84] in 1950 published a method of sequentially testing a null hypothesis and this procedure may also be applied to the selection of judges for a taste panel. The theory of the procedure needs further investigation since its properties are not well known. However, when it is applied to sequences of triangle tests, apparently satisfactory results are obtained.

Lombardi in conjunction with the present author has adapted the Rao method to use with the binomial distribution. The Rao procedure differs from that of Wald in that a limit to the testing of any one potential judge may be set. We define N to be the maximum number of tests to be given to any judge. Only one limit of ability is set. Individuals having ability greater than p_1 will be accepted and those with ability less than p_1 will be rejected. α is the probability of selecting an unacceptable judge.

One central line L is used with slope p_1 and intercept

$$h = \frac{Np_1}{\alpha}\left[I_{p_1}(n_0 - 1, N - n_0 + 1) - \alpha\right] \tag{10}$$

where $I_{p_1}(n_0 - 1, N - n_0 + 1)$ is the incomplete beta function. n_0 is the minimum number of successes in N trials which rejects p_1 as a judge's ability in favor of the alternative $p > p_1$ at the significance level α. Approximately,

$$n_0 = p_1N + \sqrt{2.72Np_1(1 - p_1)} \quad \text{if} \quad \alpha = .05$$

and

$$n_0 = p_1N + \sqrt{5.43Np_1(1 - p_1)} \quad \text{if} \quad \alpha = .01.$$

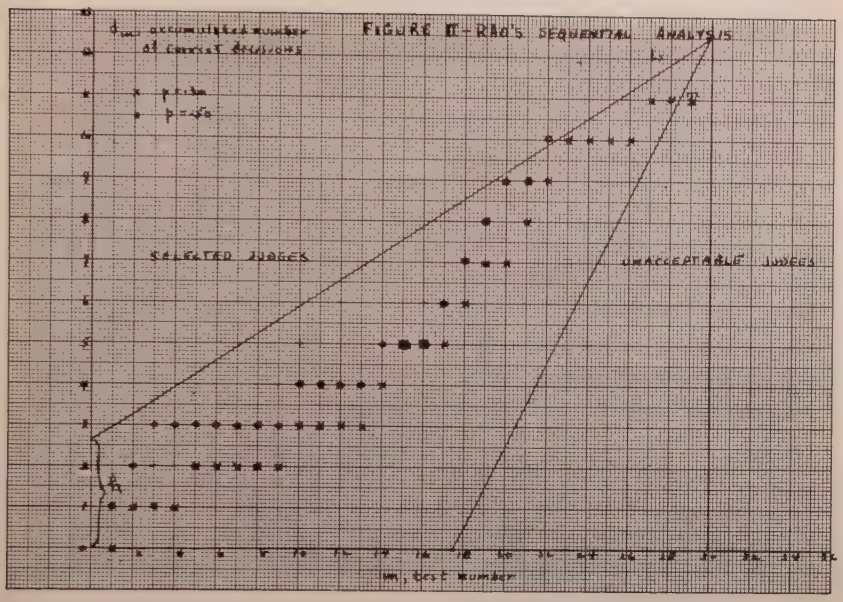
Tables of the incomplete beta function are available [70].

Let us consider an example wherein $p_1 = .33$, $N = 30$, $\alpha = .05$. Then $n_0 = 15$ and $h = 2.62$. The equation of L becomes

$$L: \quad d_m = .33m + 2.62. \tag{11}$$

where d_m is the accumulated number of correct decisions and m is the number of tests.

Two judges with abilities .3 and .5 have been simulated using tables of random numbers. The procedure is illustrated in Figure II. An additional line T named the truncation line has been drawn. T is such that if the sequence of trials leads to a point below T it will be impossible to accept the judge within the specified maximum number



of tests. One should then, if this happens, stop testing at such a point. T is drawn with slope 1 through the intersection of L and the vertical line $m = N$.

Both methods of sequential selection of judges presented have features which are desirable. The computation for Rao's procedure is simpler than for that of Wald if tables of the incomplete beta function are available. In the Rao method in practice one would usually choose p_1 considerably larger than in the illustrative example.

III. THE DESIGN OF TASTE TESTING EXPERIMENTS

3.1 *Problems of Scaling and Scoring.*

In most experimentation the units of measurement are clearly defined and follow naturally from the formulation of the problem. In a majority of taste testing problems it is impossible to avoid using scores or orderings based on the purely subjective opinions of judges or panels of judges. The statistical problems peculiar to this field of experimentation arise out of the limitations and uncertainties of human behavior. Some of the difficulties introduced are non-uniformity of scoring procedures, lack of consistency on the part of an individual judge, lack of agreement between a judge and what may be regarded as a "true criterion", doubt as to the appropriateness of analyses of variance on scores or orderings, and limitations of numbers of observations due to taste fatigue.

Scoring scales in taste testing are usually set up arbitrarily. A range of values symmetrically placed about an origin are used to measure variation from poor to excellent as related to some food attribute. The choice of scoring scale has been discussed by Hopkins [78]. When a scale has been specified, there is no guarantee that all members of a panel will utilize it in the same way. Panel members may spread their scores over the complete range or may use only a segment of the scale. Too, they may record scores falling in similar size ranges but differ in the location of their average scores. To offset these difficulties standards are sometimes used although all too often only one is specified. The standards in effect have been given a prespecified score and test items are scored in comparison with the standards. The following quotation by Harrison and Elder [82] is pertinent.

"A numerical scale for scoring often proves to be a 'rubber yardstick' unless the precaution is taken to include more than one fixed standard at predetermined points on the scale for the orientation of the taster. Repeatedly we have observed that a wide average score difference, observed between two samples when presented by themselves, will shrink when a third higher quality sample is presented along with them. This phenomenon is familiar to the psychologist who recognizes it as 'adaption'."

Alternate to using discrete scores, Baten [72, 73] has used an interesting continuous line scoring system ranging from very poor to very good. He indicated in his earlier paper that increased accuracy was obtained by the method but that it proved unpopular with the judges. Standards would again be required to obtain uniformity of scoring.

Most scoring systems lead to some doubt as to the validity of analysis of variance techniques and it is not clear how serious this may be. The basic assumptions of analysis of variance [7] are

(i) Observations are independently distributed,
(ii) Observations come from a normal population,
(iii) Error variances are homogeneous,
(iv) Treatment and environmental effects are additive.

Taste fatigue may introduce departures from (i) and (iii). A discrete scoring method without the use of standards usually leads to departures from (iii) and, with the effect of 'adaption', from (iv). The discreteness of a scale always leads to violation of (ii) although this may not be too serious.

Taste fatigue and the necessity of including at least two standards in a tasting experiment require either a reduction in the number of treatments to be compared or the use of incomplete block designs.

Difficulties in the establishment of suitable scales suggest more extensive use of ranking methods than has been made to date.

3.2 *Experimental Designs.*

Designs which increasingly control sources of variation and permit flexibility in experimental procedures are the randomized block, the latin square, the split-plot, the factorials, and the more complex incomplete block designs and lattices. It is a good rule, applicable in taste testing as well as elsewhere, to use the simplest possible design that meets the need of the experiment. However, the incomplete block designs are particularly appropriate to taste testing. The number of samples tasted in one sitting is usually limited sharply and the use of small blocks of treatments is of great assistance. It should also be noted that this eliminates the need for the judge to have longer term memory for he need only be consistent on his level of judgment within the incomplete block unit. An excellent reference to experimental designs by Cochran and Cox [19] is available.

Latinized Rectangular Lattices developed by Harshbarger and Davis [25] are of some particular interest. In most incomplete block designs the analysis separates out the effects of replications, treatments and blocks and interaction effects are not available. The latinized rectangular lattice introduces something of the advantage of the latin square. The available effects are sets, rows, interaction of rows by sets, blocks, and treatments. Then it is possible to associate judges with sets and days with rows and obtain some indication of judge behaviour from the interaction effect. This seems to constitute an advantage over other incomplete block designs. The design is available for treatment numbers which are products of two consecutive integers, $k(k-1)$.

It is believed that the analysis of variance may be used without serious error in some taste testing experiments but that in cases of doubt ranking methods should be substituted when they are available.

3.3 *Ranking techniques*

The usual criticism of ranking methods stems from a supposed loss in efficiency. When quantitative judgments can be obtained, the magnitude of differences is obscured by the use of ranks. On the other hand, when treatment differences are very small and difficult to detect, it would appear reasonable to simplify the procedure for the judge and use a ranking technique. In many cases of the latter sort the use of a scoring scale or the continuous line may give the appearance of a precision of judgment which does not in fact exist. Again, the

application of rank order methods is usually computationally simple and they may be often preferred on this ground alone. An interesting discussion on this subject was recently presented by White [58].

The modern development of rank tests has been largely limited to tests of two treatments. Such tests have been considered by Mann and Whitney [53], Wilcoxon [59, 60], Wald and Wolfowitz [56] and others. These tests could often be conveniently used to replace the *t*-test in sensory testing. Terry [55] has developed a test of this type which has the feature of being the most powerful rank order test in situations where a *t*-test would have been appropriate if quantitative measures could have been taken. His test depends on order statistics as obtained by transforming ranks using Table XX, *Scores for Ordinal (or Ranked) Data* of Fisher and Yates [46].

When ranks are employed in the analysis of variance, it has become common in taste-testing problems to transform these ranks using Table XX. Bliss [10] has supported the procedure and other references may be found in the bibliography. That the procedure is most powerful in that case considered by Terry may suggest confidence in its more general use.

Rank order methods analogous to the analysis of variance have been considered. The methodology is not complete and usually involves approximations since the computation of exact probability tables is often an exhaustive process. Kendall [52], Friedman [47] and Kendall and Babington Smith [51] have considered tests of agreement or concordance with ranked data and these methods are applicable to taste testing. Mood [54] (c.f. Chap. 16) has also contributed useful rank order methods developed from a slightly different viewpoint.

3.4 *Paired Comparisons.*

The method of paired comparisons may be regarded as a special rank order technique. It is a method long used in psychological experimentation and one that is well adapted to sensory difference testing. Only two treatments need be considered at one time and qualitative decisions alone are required. The design becomes somewhat cumbersome if many items are compared but hidden replication offsets some of that difficulty.

Two somewhat comparable methods of analysis have been presented by Thurstone [38] and by Terry with the present author [28, 37]. The mathematical formulations of the models are apparently different but may be related. Mosteller [33] has summarized Thurstone's model and listed the following underlying principles [c].

[c] Actually Mosteller lists six principles. The remaining two relate to the method of experimentation and the purpose of the analysis.

(1) There is a set of stimuli which can be located on a subjective continuum or sensation scale.

(2) Each stimulus, when presented to an individual, gives rise to a sensation in the individual.

(3) The distribution of sensations from a particular stimulus for a population of individuals is normal.

(4) It is possible for paired sensations to be correlated. The model may in a sense be summarized by writing

$$\pi_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-(S_i - S_j)}^{\infty} e^{-\frac{1}{2}y^2}\, dy \tag{12}$$

where S_i and S_j are the "true" treatment locations on the sensation continuum and π_{ij} is the probability that treatment i be rated above treatment j.

In the second method of analysis the mathematical model is formulated as follows:

(1) t treatments in an experiment using paired comparisons have true ratings $\pi_1, \cdots, \pi_t$ $(\pi_i \geq 0)$.

(2) Observations on pairs of treatments are independent in probability.

(3) When treatment i is compared with treatment j, the probability π_{ij} that treatment i be rated above treatment j is $\pi_i/(\pi_i + \pi_j)$. (This further specifies the nature of the true ratings).

If we redefine

$$\pi_{ij} = \frac{1}{4} \int_{-(\log_e \pi_i - \log_e \pi_j)}^{\infty} \operatorname{sech}^2 \frac{y}{2}\, dy, \tag{13}$$

it is easy to show that then $\pi_{ij} = \pi_i/(\pi_i + \pi_j)$. Thus the substitution of the 'squared hyperbolic secant' density for the normal density of Thurstone's model yields the second method of analysis[d]. The squared hyperbolic secant density is very similar to the normal. The substitution in Thurstone's model is a sufficient condition for the application of the model developed by Terry and the author. It would not appear to be necessary. Values $\log_e \pi_i$ correspond to values S_i on a subjective continuum. Methods developed for the estimation of these sets of parameters differ.

In the second procedure estimates p_i of π_i are obtained by the method of maximum likelihood. When some one treatment is always rated above all others, that treatment obtains a rating $p = 1$ while the others have relative ratings zero. It then appears that one good

dThe author is indebted to Professor R. A. Fisher for a suggestion on this point.

treatment obscures the differences among the others. This is in accordance with the idea of "adaption". Actually secondary ratings may be obtained for the remaining treatments by considering the subexperiment consisting of those $(t - 1)$ treatments. These secondary ratings were actually used in the evaluation of test statistics to distinguish between experimental results which had the one 'perfect' treatment but differed otherwise. This is a point which has not heretofore been exhibited.

The second test formulation has considerable flexibility and for the detection of treatment differences the results of several judges may be combined without the requirement of uniformity of ranking judgments over the judges. Fairly extensive tables are available [28] for the easy application of this method and further computation is in progress.

Two other methods of paired comparisons are of interest. Kendall and Babington Smith [31] proposed a method which is a combinatorial type test. They form a coefficient of agreement which essentially measures discrepancies from perfect agreement among judges and a coefficient of consistency for a single judge measured in terms of "circular triads".

Scheffé [36] has developed a method of paired comparisons which differs from the others in that it uses a scoring method and the analysis of variance. The method has the feature that the effect of order of presentation of paired samples to the judges is taken into account. This method seems admirably suited to consumer preference studies wherein a considerable time lag may occur between the testing of the two samples of a pair.

IV. DISCUSSION AND SUMMARY

In discussing a subject of the scope of the title of this paper certain sacrifices must be made. For completeness some topics have been included and discussed in a very superficial manner. Others may seem to have received more attention than is warranted. It was felt that the sections on sequential analysis were important to bring that phase of statistics to the attention of those interested in food and color testing by subjective appraisals. Emphasis was placed on the method of paired comparisons since some doubts and misunderstanding regarding the two models shown have arisen.

Most of the remarks have dealt with statistical methods for taste testing for differences. The subject of quality evaluation is a difficult one and one which requires considerable study. The author has been interested in the application of discriminant function techniques to the establishment of weights for the scores of various attributes in

grading. This would appear possible if reliable grade determinations could be obtained independent of the present systems of weights. It is not clear how this could be managed and any practical applications of the technique would seem to involve somewhat circular arguments.

In conclusion the author would like to acknowledge the assistance of M. E. Terry, Boyd Harshbarger, Lyle L. Davis and D. B. Duncan in the studies of statistical methods for taste testing undertaken at the Virginia Agricultural Experiment Station.

V. CLASSIFIED BIBLIOGRAPHY

ANALYSIS OF VARIANCE

1. Bartlett, M. S. The use of transformations, *Biometrics* 3, 39, 1947.
2. Cochran, W. G. Problems arising in the analysis of a series of similar experiments, *Jour. of the Royal Statistical Society*, Supplement 4, 102, 1937.
3. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied, *Biometrics* 3, 22, 1947.
4. Cox, G. M. Statistics as a tool for research, *Journal of Home Economics*, 36, 575, 1944.
5. Crump, S. L. The estimation of variance components in analysis of variance, *Biometrics* 2, 7, 1946.
6. Curtiss, J. H. On transformations used in the analysis of variance, *Annals of Math. Stat.* 14, 107, 1943.

6a. Duncan, D. B. A significance test for differences between ranked treatments in an analysis of variance, *Va. Jour. of Sci.*, 2, 171, 1951.

6b. Duncan, D. B. On the properties of the multiple comparisons test, *Va. Jour. of Sci.*, 3, 49, 1952.

7. Eisenhart, C. The assumptions underlying the analysis of variance, *Biometrics* 3, 1, 1947.
8. Snedecor, G. W. *Statistical Methods*, Iowa State College Press, Ames, 1946.
9. Tukey, J. W. Comparing means in the analysis of variance, *Biometrics* 5, 99, 1949.

CONSUMER PREFERENCES

10. Bliss, C. I., Anderson, E. D. and Marland, R. E. A technique for testing consumer preferences with special reference to the constituents of ice cream, *Storrs Agricultural Experiment Station Bulletin* 251, 1943.
11. Bogert, J. L. A method of consumer product testing, *Food Industries* 13, 47, 1941.
12. Garnatz, G. But what do the consumers say? *Food Industries* 22, 1333, 1950.
13. Nair, J. H. Mass taste panels, *Food Tech.* 3, 131, 1949.

DISCRIMINANT FUNCTION ANALYSIS

14. Baten, W. D. The use of discriminating functions in comparing judges' scores concerning potatoes. *J.A.S.A.* 40, 223, 1945.
15. Brown, G. W. Discriminant functions. *Annals of Math. Stat.* 18, 514, 1947.
16. Cochran, W. G. and Bliss, C. I. Discriminant functions with covariance. *Annals of Math. Stat.* 19, 151, 1948.

17. Johnson, P. O. The quantification of data in discriminant analysis. *J.A.S.A.* 45, 65, 1950.

EXPERIMENTAL DESIGNS

18. Bose, R. C. Partially balanced incomplete block designs with two associate classes involving only two replications. *Calcutta Statistical Association Bulletin* 3, 120, 1951.
19. Cochran, W. G. and Cox, G. M. *Experimental Designs.* John Wiley and Sons, New York, 1950.
20. Connor, W. S. On the structure of balanced incomplete block designs. *Annals of Math. Stat.* 23, 57, 1952.
21. Greenwood, M. L., Potgieter, N. and Bliss, C. I. The effect of certain pre-freezing treatments on the quality of eight varieties of cultivated high bush blueberries. *Food Research* 16, 154, 1951.
22. Hanson, H. L., Cline, L. and Lineweaver, H. Application of balanced incomplete block design to scoring of ten dried egg samples. *Food Tech.* 5, 9, 1951.
23. Harshbarger, B. Triple rectangular lattices. *Biometrics* 5, 1, 1949.
24. Harshbarger, B. Near balanced rectangular lattices. *Va. Jour. of Sci.* 2 (New Series), 13, 1951.
25. Harshbarger, B. and Davis, L. L. Latinized rectangular lattices. *Biometrics* 8, 73, 1952.
26. Hopkins, J. W. Laboratory flavor scoring—a two-factor experiment in complete and incomplete blocks. *National Research Council, Canada, Publication, Committee on Applied Mathematical Statistics* No. 7 (mimeo).
27. Shrikhandi, S. S. Designs for two-way elimination of heterogeneity. *Annals of Math. Stat.* 22, 235, 1951.

PAIRED COMPARISONS

28. Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs I. To be published in *Biometrika*, 39, 1952.
29. Guttman, L. An approach for quantifying paired comparisons and rank order. *Annals of Math. Stat.* 17, 145, 1946.
30. Hening, J. C. Flavor evaluation procedures, *New York Agricultural Experiment Station Technical Bulletin* 284, 1948.
31. Kendall, M. G. and Babington Smith, B. On the method of paired comparisons. *Biometrika* 31, 324, 1939–40.
32. Moran, P. A. P. On the method of paired comparisons, *Biometrika* 34, 363, 1947.
33. Mosteller, F. Remarks on the method of paired comparisons I. *Psychometrika* 16, 3, 1951.
34. Mosteller, F. Remarks on the method of paired comparisons II. *Psychometrika* 16, 203, 1951.
35. Mosteller, F. Remarks on the method of paired comparisons III. *Psychometrika* 16, 207, 1951.
36. Scheffé, H. On analysis of variance for paired comparisons, *J.A.S.A.* 47, 381, 1952.
37. Terry, M. E., Bradley, R. A. and Davis, L. L. New designs and techniques for organoleptic testing. *Food Tech.* 6, 250, 1952.
38. Thurstone, L. L. Psycho-physical analysis. *Am. Jour. of Psychology* 38, 368, 1927.

39. Thurstone, L. L. *An Experiment in the Prediction of Choice.* Publication of the Psychometric Laboratory, University of Chicago, No. 68, 1951.

RANK ORDER METHODS

40. Baten, W. D. Analysis of scores from sampling tests. *Biometrics* 2, 11, 1946.
41. Baten, W. D. and Trout, G. M. A critical study of the summation difference-in-rank method of determining proficiency in judging dairy products. *Biometrics* 2, 67, 1946.
42. Charley, H. Effect of baking pan material on heat penetration during baking and on quality of cakes made with fat. *Food Research* 15, 155, 1950.
43. Dixon, W. J. The criterion for testing the hypothesis that two samples are from the same population. *Annals of Math. Stat.* 11, 199, 1940.
44. Dixon, W. J. and Mood, A. M. The statistical sign test. *J.A.S.A.* 41, 557, 1946.
45. Durbin, J. Incomplete blocks in ranking experiments. *Br. Jour. of Psychology,* Stat. Section 4, 1951.
46. Fisher, R. A. and Yates, F. *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh, 1948.
47. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J.A.S.A.* 32, 675, 1937.
48. Galinat, W. C. and Everett, H. L. A technique for testing flavor in sweet corn. *Agronomy Journal* 41, 1949.
49. Greenwood, M. L. and Salerno, R. Palatability of kale in relation to cooking. *Food Research* 14, 314, 1949.
50. Hotelling, H. and Pabst, M. R. Rank correlation and tests of significance involving no assumption of normality. *Annals of Math. Stat.* 7, 29, 1936.
51. Kendall, M. G. and Babington Smith, B. The problem of m rankings. *Annals of Math. Stat.* 10, 275, 1939.
52. Kendall, M. G. *Rank Correlation Methods.* Charles Griffin and Co., London, 1948.
53. Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Math. Stat.* 18, 50, 1947.
54. Mood, A. M. *Introduction to the Theory of Statistics.* McGraw-Hill, New York, 1950.
55. Terry, M. E. Some Rank Order Tests Which Are Most Powerful Against Specific Parametric Alternatives. *Annals of Math. Stat.,* 23, 346, 1952.
56. Wald, A. and Wolfowitz, J. On whether two samples are from the same population, *Annals of Math. Stat.* 11, 147, 1940.
57. Walsh, J. E. Some significance tests for the median which are valid under very general conditions, *Annals of Math. Stat.* 20, 64, 1949.
58. White, C. The use of ranks in a test of significance for comparing two treatments, *Biometrics* 8, 33, 1952.
59. Wilcoxon, F. Individual comparisons by ranking methods, *Biometrics* 1, 80, 1945.
60. Wilcoxon, F. Probability tables for individual comparisons by ranking methods, *Biometrics* 3, 119, 1947.

REFERENCE MATERIAL

61. *Minutes of Conference on Taste Panel Procedures and Methods of Evaluation for Soy Bean Oil and Soy Bean Oil Products,* Northern Regional Research Laboratory, Peoria, Ill., 1949 (mimeo).

62. *Taste Panels, Bibliography.* Quartermaster Food and Container Institute, Chicago, Ill., (mimeo).
63. Bradley, R. A. and Duncan, D. B. *Bi-Annual Report No. 1 on Statistical Methods for Sensory Difference Testing,* (mimeo), 1950.
64. Bradley, R. A. and Terry, M. E. *Bi-Annual Report No. 2 on Statistical Methods for Sensory Difference Testing,* (mimeo), 1951.
65. Bradley, R. A. and Terry, M. E. *Bi-Annual Report No. 3 on Statistical Methods for Sensory Difference Testing,* (mimeo), 1951.
66. Bradley, R. A. and Terry, M. E. *Bi-Annual Report No. 4 on Statistical Methods for Sensory Difference Testing,* (mimeo), 1952.
67. Dawson, E. H. and Harris, B. L. Sensory methods for measuring differences in food quality. *Agricultural Information Bulletin No. 34,* U.S.D.A., 1951.
68. Gilford, J. P. *Psychometric Methods.* McGraw-Hill, New York, 1936.
69. Justin, C. and Adams, G. *Published and Processed Reports of Research in Foods.* Human Nutrition and Home Economics at the Land-Grant Institutions, Oct. 1950–Oct. 1951, Office of Experiment Stations, Agricultural Research Administration, U.S.D.A. (mimeo).
70. Pearson, K. *Tables of the Incomplete Beta-Function.* Cambridge University Press, Cambridge, England, 1934.
71. Scheffé, H. Statistical inference in the non-parametric case, *Annals of Math. Stat.* 14, 305, 1943.

SCORING TECHNIQUES

72. Baten, W. D. Organoleptic tests pertaining to apples and pears, *Food Research* 11, 84, 1946.
73. Baten, W. D. Reaction of age groups to organoleptic tests. *Food Tech.* 4, 277, 1950.
74. Boggs, M. M. and Ward, A. C. Scoring techniques for sulfited foods. *Food Tech.* 4, 282, 1950.
75. Cochran, W. G. The comparison of different scales of measurement for experimental results. *Annals of Math. Stat.* 14, 205, 1943.
76. Dove, W. F. The relative nature of human preference. *Jour. of Comparative Psychology* 35, 219, 1943.
77. Hening, J. C. Operations of a routine testing group in a small laboratory, *Food Tech.* 3, 162, 1949.
78. Hopkins, J. W. A procedure for quantifying subjective appraisals of odor, flavor, and textures of foodstuffs. *Biometrics* 6, 1, 1950.
79. Norton, H. W. Uses of scores in an exact test of significance in a discontinuous distribution. *Annals of Eugenics* 7, 349, 1937.
80. Ward, A. C. and Boggs, M. M. Comparison of scoring results for two and four samples of corn per taste session. *Food Tech.* 5, 219, 1951.

SEQUENTIAL ANALYSIS

81. Anscombe, F. J. Large-sample theory of sequential estimation. *Biometrika* 36, 455, 1949.
82. Harrison, S. and Elder, L. W. Some applications of statistics to laboratory taste testing. *Food Tech.* 4, 434, 1950.
83. Lombardi, G. J. *The Sequential Selection of Judges for Organoleptic Testing.* (Thesis) Virginia Polytechnic Institute, 1951.
84. Rao, C. R. Sequential tests of null hypotheses. *Sankhya* 10, 361, 1950.
85. Wald, A. *Sequential Analysis.* John Wiley and Sons, New York, 1947.

SOME PROBLEMS IN THE DESIGN AND STATISTICAL ANALYSIS OF TASTE TESTS.[1]

David D. Mason and E. James Koch[2]

Introduction:

Taste tests may be considered as a part of the general classification of subjective tests. For the purpose of this discussion, subjective tests are defined as those tests in which a particular characteristic or property of a material or commodity is scored or otherwise rated by an individual or individuals, with this score or rating being decided upon by judgement. These are contrasted to objective tests, whose outcome, as the name implies, is largely independent of human judgement. It is, of course, appreciated that objective tests are usually a more precise measure of a particular property than are subjective tests, and where possible and feasible, objective tests would be used. However, there are certain characteristics, such as flavor, odor, and most especially, individual preference, for which no satisfactory objective tests are available. Therefore, if one wishes to obtain information on these types of properties, he must be prepared to deal with the problems attendant to subjective tests. We feel that, by paying proper attention to the planning, design and execution of the experiment, that we have generally been able to provide satisfactory answers to the research workers' objectives, relating to these properties.

It is the purpose of this discussion to outline, from our experience, some of the major problems and stumbling blocks in the planning and executing of taste test experiments. This experience has been limited in scope in that we have been largely dealing with two types of problems; (1) determining the effect of methods of storage and handling on the flavor and other pertinent quality factors on horticultural crops, and (2) taste preferences as a quality factor in evaluation of variety or selection tests in a plant breeding program.

---

[1]Presented to Joint Symposium of Biometric Society and the American Society of Horticultural Science, Annual Meeting, Cornell University, Ithaca, New York, September 9, 1952.

[2]Biometrician and Associate Biometrician, respectively, Bureau of Plant Industry, Soils and Agricultural Engineering, U.S.D.A., Beltsville, Maryland.

Factors for Consideration in Preliminary Planning

The first and most important fact that one must have in the planning of the taste test is the answer to the question: "What use is to be made of the results?" If no good answer can be provided, it is obviously not worthwhile to run the experiment. The answer to this question determines the type of subjects to be used, the quality characteristics to be evaluated, the sample preparation, the type of ratings used, etc. Once the objectives of the experiment have been clearly stated, the greatest hurdle in the planning is over. We hesitate to mention such an obvious fact in this discussion, but since this is, in our opinion, perhaps the most important reason for the failure of taste tests, we feel that it should be emphasized.

The other steps in planning are to be considered briefly. They are not necessarily mentioned in the order of their importance. Failure to adequately consider any one of the factors will mean failure, or at least serious impairment of the experiment.

Type of Judges Panel: Whether to use an expert panel, or a consumer preference type panel depends on the objectives of the experiment. There is no particular reason to discuss the relative merits of these two types of panels, for they serve quite different purposes. It is obviously impossible to discuss all the methods and procedures for panel selection; the problem is reviewed in (1) (2).

One of the common problems in either type panel is proper panel instruction. It must be remembered that the panel, particularly the consumer preference type, can be given too much information about the samples, as well as too little. The method of scoring or rating must be thoroughly understood by all panels. It is a point easily overlooked in preference panels.

Selection of Rating System and The Score Card: The system used in horticultural work at B.P.I.S. & A.E. is essentially that reported by Morrow, Darrow and Rigney (4) in a paper entitled, "A Rating System for the Evaluation of Horticultural Material." A scoring system with a range from 1 to 10 is used. The score of 1 indicates the poorest, and a score of 10, the best rating for a particular character. In an attempt to separate differences more clearly in the mind of the individual doing the rating, we have considered ratings from 1 to 5 to be below the limits for commercial usage, and the ratings of 6 to 10 to be within the range for commercial usage. In the instructions we further specify that ratings of 9 to 10 are of excellent quality.

This scoring system had been used previously by research workers in the Bureau for disease ratings, etc., and for this purpose had been

found to be a very satisfactory system. We have continued to use this system because we have found that it works satisfactorily for most taste tests and felt that a uniform system throughout the Bureau would be desirable. With a uniform system, less panel instruction needs to be given at the beginning of each test.

Other scoring systems have also been described. Hopkins (3) has reported a detailed investigation on the use of an 11 point scoring system, with which he has shown a close and consistent relationship between scores and concentration of certain components. We feel that a 1 to 10 scoring system, once it has been used by the tasters, is just as satisfactory as the -5 to $+5$ scoring system suggested by Hopkins. At the same time, the 1 to 10 scale avoids negative values which are hazardous in both recording and computing.

We have two reasons for choosing a scoring system in preference to a ranking system where the treatments are ranked according to a preference for a given product. The first reason is that a scoring system permits ties, whereas a ranking system forces judgement even though the taster can detect no definite difference. The second reason is that a scoring system permits the spread of treatments to be influenced by the magnitude of the differences found. If there are very distinct differences between two samples, a taster can score them 10 to 1 whereas, in ranking they must be ranked 1 and 2 respectively.

It is realized that there is danger in comparing scores of different experiments on similar material, because the rating of a particular sample is influenced by the relative level of the samples with which it is tested. Yet, because of the previously stated standards which have been set up for our scores, it does provide for a uniform system of reporting.

In making up the score card, we have found that the simpler the card, the better. Briefly, we use 3×5 cards, with the number of rows (for samples) corresponding to the number of items in the block (number of samples to be tested at one time), and the number of columns corresponding to the number of characteristics to be rated. There is always a tendency to try to obtain too much information about a particular sample. Again, this goes back to the objectives of the experiment. If no clearcut need for information about a particular character can be seen, it should be omitted. In our experience, when the number of characteristics to be rated in a test goes over three, the efficiency of the judges drops sharply. An element of confusion seems to be introduced into the judges' mind when he must consider a number of characters.

Selection and Preparation of the Material Included in the Test:

In the preparation of material for the actual taste test, it is very easy to overlook certain important factors that may introduce unwanted bias into the results. For example, the factor of ripeness at time of harvest is a very important factor in testing either fresh or processed fruits. It is extremely easy to have the scores reflect the degree of ripeness rather than the real quality of the fruit at the same stage of ripeness. Also, it is obvious that the uniformity of the handling of the material in processing is very important in order to avoid excessive and unwanted variation in the results. For example, the amount of sugar or salt, or other agent added in the preparation (unless it is a variable in the test) should be controlled very carefully.

Also, it goes without saying that the preparation of the actual samples for submission to the judges must be carefully supervised. For example, if one is testing frozen food products, it is extremely important that all the treatments be brought to the same temperature, texture, color etc. before they are judged.

Judging Conditions:

For sensory difference testing, where expert panels are used, and high precision is essential, optimum and uniform test room conditions are necessary. For preference testing, such conditions are desirable, but perhaps not as essential. Many horticulturalists are faced with the problem of lack of adequate facilities to satisfactorily handle a large panel. However, judging under poor aesthetic conditions, or where judges may converse and observe each other's scores and opinions, is highly undesirable.

Lately, we have been using a procedure that may, at first, seem objectionable. But it must be considered in the light of a number of factors. The procedure is to supply the samples to the judges in their offices or laboratories. We first started this procedure in testing peanut butter for off-flavor possibly caused by BHC. After some experience, it was apparent that there should be some lapse of time between samples, because often, there was a delayed taste reaction to the off-flavor caused by BHC. Since our panel is made up largely of professional personnel for whom it is inconvenient to spend an excessive amount of time away from their regular duties, and since we had a relatively large number of samples to test, the procedure of taking the samples to the judges was initiated. It appeared to work so satisfactorily in this case, that we tried it for other products with good results.

Designs for Taste Testing Experiments:

The number of samples (varieties or treatments) for which we have set up designs have ranged from 8 to 64. Since our experience, and that of others has indicated that a maximum of 8 samples can be adequately evaluated by a judge at any one time, we have normally used some form of incomplete block design. As a general rule, we try to keep the number of samples judged at any one time down to 6 or below. Specifically, the incomplete block designs used satisfactorily have been the balanced incomplete blocks, balanced lattices, rectangular lattices, and the latinized rectangular lattice.

An example of the layout of a particular design might be given. Twenty-four treatment combinations of different picking conditions and storage conditions of one variety of peaches were packed and frozen under uniform processing. The objective was to determine the effect of the picking and storage conditions on texture, and flavor of the frozen product.

For this number of treatments, we would add a standard sample to bring the number of entries to 25. Thus, we could use a 5×5 lattice. From previous experience with this type of material, and in order to get a representative panel, it was decided that 3 repetitions of the 5×5 balanced lattice would be adequate. We would thus have 18 replications. Using judges for replications, we would then have 18 judges. After going through the necessary procedures of randomization, etc., each judge would evaluate 5 samples at a time. He would be engaged in the test for five days.

In this design, then, we have control over the major factors of variability. The variation between judges may be removed as replications. The variation of the same judge from day-to-day may be removed as blocks, as well as any other day-to-day variation.

The coefficient of variation in such experiments as these usually are from 15-25%. By having from 15 to 20 replications (judges), we usually have "Least Significant Differences" (L.S.D.'s) of about 1.0 unit, at the 5% significance level. This level of precision seems to be adequate to meet the objectives of most experimenters.

With tests where the primary objective is to compare a treatment with some standard or check, or where the material is of such a nature that only 2 or 3 samples at the most can be accurately evaluated at one time, the paired comparison, duo-trio test or triangle test may be used to advantage.

A brief case history of such a situation is given by our experience with tests with peanut butter samples. In a cooperative project between

the states concerned and the Bureau of Entomology and Plant Quarantine, the Bureau of Human Nutrition and Home Economics and our Bureau, an endeavor was made to determine the extent to which the use of BHC, and some other insecticides on cotton, preceding a crop of peanuts, had on imparting a detectable off-flavor to the peanut in the form of peanut butter. After some preliminary work, it was established that the judges could not handle more than one treated sample at a time to the lingering effects of a very strong dosage. Since the main objective was to detect off-flavor, the comparison with a standard or untreated sample was the most important to make. Therefore, a paired comparison test was used, in which the judge was submitted two samples identified only by code, but one of which was a check, and one a test sample. He was required to rate the samples only on flavor.

This test worked out fairly well, but it was noted that the judges were becoming less discriminating after testing several sets of samples (two sets of two samples each were tasted each day by each judge—one in the morning, and one in the afternoon.) A panel of 20 judges was used. The general complaint of the judges was that they had lost confidence in their ability to discriminate.

This past year, in a similar test, a modification of the duo-trio test was used. Again, a panel of 20 judges was used. Each judge was given three samples, (1) a marked standard, (2) an unmarked standard, and (3) an unknown. He was requested to identify the unmarked standard. He was also asked to score the samples. This design appeared to work out better, in that it gave the judges a constant reference standard for purposes of orientation.

The Bureau of Human Nutrition and Home Economics ran a parallel test on the same samples, using a smaller, trained, expert panel selected to detect differences. The final results from the two panels have been quite similar.

In summary, the problems of the design of experiments can be met by employing the same basic principles of design that apply to other fields of biological research. The incomplete block designs are almost a necessity to cope with the human fatigue factor and still maintain continuity between all treatment comparisons in a given experiment. The alternative of subdividing the treatments of an experiment into groups of smaller experiments in order to meet the fatigue problem does not allow the objectives to be satisfied in most cases.

Statistical Analysis of Taste Test Data:

There seems to be two principal concerns in the analysis of scores

from taste tests. The first problem is whether the assumptions implicit in the analysis of variance have been met. Hopkins (3) has discussed the problem rather thoroughly, and points out that the variances and means are not independent when the full range of the scoring system is used, the distribution being analogous with binomial variation. He suggests that an angular transformation would be expected to render the means and variances independent. Hopkins work also demonstrated, and he concluded that in most cases, where the values are near the center of the range, the variances and means may be assumed to be independent. This is the range that most workers are interested in. In fact, if some variety or treatment is abnormally low, or abnormally high, the conclusions are usually self-evident, and require no precise testing.

Our practice has generally been to use the raw scores for analysis, without transformation. We endeavor to keep a watch for abnormalities such that might seriously disturb the tests of significance. With a large panel, it appears, barring extremes, that the data come about as near to normal distribution as do many other types of biological data that we work with.

We will welcome the development of the ranking methods which will allow significance tests without the assumptions necessary in the analysis of variance. In the meantime the principal problems may be kept under some control by careful observation of score distributions, and by adequate judge instruction to insure their using the scale properly.

The second difficult problem is that of, having found a significant F test, and thereby rejecting the treatment homogeneity hypothesis, of determining which of the differences among the means are real. In many cases, the plan of the treatments allows a valid comparison (factorial experiments, for example). However, the case is common, such as in variety tests, where there is no objective basis for comparison other than the size of the means themselves. Tukey (4) and Duncan (5) are among the latest contributors to a method for the solution of the problem. However, these methods are difficult for the average applied research worker to understand and execute. We cannot expect their general acceptance except at institutions where a professional statistical staff with computing laboratory facilities are available to do the work. Meanwhile, the Least Significant Difference (L.S.D.) appears to be a very useful statistic, in spite of its frequent misuse and malignment. Their judicious use will allow the research worker to select his best performing varieties or treatments with some confidence, although the degree of confidence may not be accurately indicated. This is usually the most important function of the test.

Summary:

We have tried to approach the problem of taste test experiments with the same plan of attack used in planning and executing other experiments with biological material. In general, if the objectives are carefully prepared, and other phases of the experiment properly executed, the experimental design and statistical analysis offer problems that appear to vary in complexity, rather than in nature, from problems with other experimental material.

REFERENCES

1. Dawson, Elsie, H. et al. Sensory Methods for Measuring Differences in Food Quality. *Agriculture Information Bulletin No. 34*, U.S. Department of Agriculture, 1951.
2. Statistical Laboratory, Virginia Polytechnic Institute. Statistical Methods for Sensory Difference Tests of Food Quality. *Bi-Annual Reports* 1, 2, 3 and 4, December 1950 to June 1952.
3. Hopkins, J. W. A Procedure for Quantifying Subjective Appraisals of Odor, Flavor and Texture of Food Stuff. *Biometrics*, Vol. 6: 1–16, 1950.
4. Morrow, E. B., G. M. Darrow and J. A. Rigney. A Rating System for the Evaluation of Horticultural Material. *Proceedings of the American Society for Horticultural Science*, Vol. 53: 276–280. 1949.

ON THE UNIQUENESS OF THE LINE OF ORGANIC CORRELATION[1]

WILLIAM H. KRUSKAL

University of Chicago

Summary. A problem sometimes arising in biometric work is the representation of a multivariate distribution by a single straight line. This paper discusses the meaning and uniqueness properties of one such method of representation which has been suggested by several writers and which may be called the *diagonal line of organic correlation.* This line passes through the mean of the distribution, has its direction numbers proportional in absolute value to the standard deviations, and has the signs of its direction numbers determined by the signs of the covariances. Sampling problems are not discussed here, but only methods of representation when the population is completely known. It is shown that under reasonable assumptions the diagonal line of organic correlation is

1. The unique line among those based on first and second moments which transforms properly under translation, change of scale, and omission of coordinates, and which in addition provides the proper directions of association;

2. The unique line maximizing the probability of correct prediction, when prediction is based on a line in accord with a scheme (described precisely further on) which supposes that all possible predictions of one component based on another are equally frequent, and that the predictions are intervals whose lengths are certain multiples of the standard deviations.

Uniqueness property 1 does not depend on normality, but 2 does. Two further geometrical interpretations of the diagonal line of organic correlation are discussed; the first (Jones [4]) as a diagonal of a hyperrectangle circumscribing an ellipsoid of concentration, and the second (Teissier [3]) as a line minimizing the expected value of the area of a certain right triangle in each bivariate marginal distribution. Finally a characterization of the diagonal line of organic correlation is given (Greenall [7]) in terms of the quantiles of the marginal distributions.

[1]Work leading to this paper was sponsored in part by the Office of Naval Research.

Introduction. In biological studies dealing with two or more quantitative characteristics of the individuals of a species or other grouping, the biologist may wish to represent the joint distribution of the characteristics by a single straight line. Let us agree to call such a line a line of organic correlation.[2] Apparently there are two major motivations for working with this kind of representation: first, the desire for a concise (although, of course, incomplete) description of the distribution in order to indicate the general relative trends of the characteristics; and, second, the prospective use of such a line as a predicting mechanism. Typical quantitative characteristics considered are: skull length, logarithm of antler length, cube root of body weight, square of eye-socket diameter, etc. When the quantitative characteristics are taken as the logarithms of directly measured quantities, this portion of biometry has been called allometry, and the line of organic correlation has been called the allometric line.

The problems considered in this field may be divided into three groups: (1) Which line should be used if the population is known? (2) When sampling from the population, how to estimate and perform significance tests on the line if errors of measurement are negligible? (3) How to estimate and perform significance tests on the line if errors of measurement are substantial? Problems (1) and (2) have been considered in a recent paper by Kermack and Haldane [1] who give references leading back to earlier literature. Problem (3) is extremely difficult in general because of the question of identification (see, e.g., [2]), but fortunately in many biological studies errors of measurement are quite small compared with inherent variability in the populations. The present note deals only with problem (1).

There are various ways in which one might decide which line to use if the population were known, and such decision is an essential prerequisite to consideration of problems (2) and (3). This paper emphasizes two approaches. In the first the main requirements are that the line transforms properly under linear transformations (e.g. computational codings) in each variable separately, that it transforms properly under omission of variables, and that it indicates proper directions of association. It is then shown that only the diagonal line of organic correlation satisfies the requirements.

In the second approach, a possible predictive use of a line of organic correlation is constructed. This use is somewhat artificial, but not—I hope—atypical of actual uses. A normality assumption is made, and it is then shown that the diagonal line of organic correlation is the best line to use in carrying out the constructed predictive procedure.

[2]Although the term "organic correlation" is used freely in this paper to describe representing lines, there is nothing about the material discussed here which is really unique to biometry.

Preliminaries. In order to gain compactness without loss of precision, I shall use vector and matrix notation where appropriate; however, no properties of matrices will be used beyond the definition of multiplication. Vectors will be horizontal.

The quantitative characteristics of biological interest are considered as the p components of a p-dimensional random (vector) variable $X = \{X_1 X_2 \cdots X_p\}$. We shall make the following general assumption:

Assumption 1. X has a mean vector $\mu = \{\mu_1 \cdots \mu_p\}$ and a non-singular covariance matrix $\Sigma = \{\sigma_{ij}\}$ $(i, j, = 1, 2, \cdots, p)$.

Thus μ_i is the expected value of X_i, σ_{ii} the variance of X_i, and σ_{ij} the covariance of X_i and X_j. The assumption of non-singularity means of course that the variance of $\Sigma a_i X_i$ is not zero unless all the a_i's are zero. This non-singularity requirement may at times be weakened at the expense of greater complexity in the statements of results, and this point will be briefly discussed later. Note that no assumption of normality has yet been made.

A second general assumption is

Assumption 2. Either all covariances of Σ are positive, or else they can all be made positive by replacing some X_i's with their negatives.

In other words, no two components of X are uncorrelated, and the sign of $\sigma_{ij}\sigma_{ik}$ is the same as that of σ_{jk}. The reason for this assumption is that we wish to exclude cases such as that of positive association between X_1 and X_2, positive association between X_2 and X_3, but negative association between X_1 and X_3. In such a case it would, I believe, be fatuous to attempt to describe the distribution by a single straight line at all; more precisely, it would then not be possible in general to satisfy Criterion 4 (see ahead). Fortunately, many biological studies for which representation by a straight line is considered do satisfy Assumption 2; and in particular when $p = 2$ Assumption 2 must be satisfied unless X_1 and X_2 are uncorrelated.

The *diagonal line of organic correlation* is defined as that line passing through the point μ and having direction numbers $\lambda_i = \text{sgn}(\sigma_{i1})\sqrt{\sigma_{ii}}$. The function $\text{sgn}(y)$ is 1, 0, -1 as y is positive, zero, or negative. In other words, the diagonal line of organic correlation passes through the mean, has slopes proportional in absolute value to the standard deviations, and has the signs of its slopes so determined by the signs of the covariances that each pair of slopes contains like or unlike signs as the corresponding variables are positively or negatively correlated. The suggestion of the term "diagonal" to characterize this line comes from the econometric literature (for example [6]) and arises from a geometrical point of view to be mentioned later.

It seems convenient to represent a line by the notation (α, λ), where $\alpha = \{\alpha_1, \cdots, \alpha_p\}$ is a point through which it passes and $\lambda =$

$\{\lambda_1 \cdots \lambda_p\}$ is a set of direction numbers. We pay for this convenient notation by its lack of uniqueness; thus two lines (α, λ) and $(\bar{\alpha}, \bar{\lambda})$ are the same if and only if there exist numbers s and t $(t \neq 0)$ such that

$$\bar{\alpha} = \alpha + s\lambda, \qquad \bar{\lambda} = t\lambda.$$

When we write $(\alpha, \lambda) = (\bar{\alpha}, \bar{\lambda})$, it will be meant in this sense.

If the line (α, λ) represents the distribution well, then the ratio λ_j/λ_i shows roughly the amount of change in characteristic j as characteristic i increases one unit.

Transformation Criteria. We shall now state the transformation criteria to be imposed on a line (α, λ) in order that it be a suitable candidate as a line of organic correlation.

Criterion 1. *Dependence on first and second moments only.* α and λ depend only on μ and Σ. This functional dependence will be indicated henceforth by the notation: $\alpha(\mu, \Sigma)$, $\lambda(\mu, \Sigma)$.

Criterion 2. *Proper behavior under translation.* If X_i be replaced by $X_i + a_i$, then α_i must be replaced by $\alpha_i + a_i$ and λ must be unchanged; i.e., for any $a = (a_1 \cdots a_p)$

$$\alpha(\mu + a, \Sigma) = \alpha(\mu, \Sigma) + a$$

$$\lambda(\mu + a, \Sigma) = \lambda(\mu, \Sigma).$$

Criterion 3. *Proper behavior under change of scale.* If X_i be replaced by $b_iX_i (b_i \neq 0)$, then α_i must be replaced by $b_i\alpha_i$, and λ_i must be replaced by $b_i\lambda_i$. Criteria 2 and 3 together require proper behavior of the line under linear transformations on each variable separately such as the transformations used in computational coding. The two criteria could of course have been expressed as a single criterion.

Criterion 4. *Proper direction of association.* No λ_i is 0. If σ_{ij} is positive, then λ_i and λ_j must have the same sign; if σ_{ij} is negative, then λ_i and λ_j must have different signs. In other words, $\lambda_i\sigma_{ij}\lambda_j > 0$ $(i, j = 1, \cdots, p)$.

Criterion 4 can only be fulfilled by virtue of Assumption 2. A consequence of this criterion is that the line to be derived will depend on the signs of the σ_{ij}'s, and this may introduce difficulties (not mentioned in [1]) in the estimation problem, particularly if some of the correlation coefficients are small in magnitude and of unknown sign.

Uniqueness Under the Transformation Criteria, Bivariate Case. It is convenient first to show the uniqueness of the diagonal line of

organic correlation under the transformation criteria in the bivariate case. Recall that in this case Assumption 2 simply means that $\sigma_{12} \neq 0$.

Theorem 1.[3] When $p = 2$ and Assumptions 1 and 2 hold, the unique line satisfying Criteria 1 through 4 is the diagonal line of organic correlation.

Proof. In the following manipulations s and t (possibly with subscripts) will denote real variables in statements formally preceded by "for some s" or "for some $t \neq 0$." However, this initial quantifying clause will be omitted for simplicity.

By Criterion 2

$$\alpha(0, \Sigma) = \alpha(\mu - \mu, \Sigma) = \alpha(\mu, \Sigma) - \mu + s_1\lambda(\mu, \Sigma). \tag{1}$$

Similarly

$$\alpha(0, \Sigma) = \alpha(-\mu, \Sigma) + \mu + s_2\lambda(\mu, \Sigma). \tag{2}$$

By Criterion 3 with $b_1 = b_2 = -1$

$$\alpha(-\mu, \Sigma) = -\alpha(\mu, \Sigma) + s_3\lambda(\mu, \Sigma). \tag{3}$$

Add (1) and (2), and simplify by means of (3)

$$\alpha(0, \Sigma) = s_4\lambda(\mu, \Sigma)$$

whence by (1)

$$\alpha(\mu, \Sigma) = \mu + s\lambda(\mu, \Sigma). \tag{4}$$

This completes the first part of the proof.

Next note that $\lambda(\mu, \Sigma)$ cannot depend on μ, for by Criterion 2

$$\lambda(\mu, \Sigma) = \lambda(0 + \mu, \Sigma) = t\lambda(0, \Sigma)$$

and we may write simply $\lambda(\Sigma) = \lambda(\sigma_{11}, \sigma_{22}, \sigma_{12})$. By Criterion 3, with $b_i = \sqrt{\sigma_{ii}}$ $(i = 1, 2)$

$$\lambda_i(\sigma_{11}, \sigma_{22}, \sigma_{12}) = t\lambda_i(1, 1, \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}})\sqrt{\sigma_{ii}}\,.$$

So we may confine ourselves to the case $\sigma_{11} = \sigma_{22} = 1$. Since interchanging coordinates does not change such a Σ, it follows that $\lambda_1/\lambda_2 = \lambda_2/\lambda_1$, or $\lambda_2 = \pm\lambda_1$. Criterion 4 selects as sign $\operatorname{sgn}(\sigma_{12})$. Thus

$$\lambda_i(\sigma_{11}, \sigma_{22}, \sigma_{12}) = t \operatorname{sgn}(\sigma_{1i})\sqrt{\sigma_{ii}}\,. \tag{5}$$

But (4) and (5) define exactly the diagonal line of organic correlation.

[3] This theorem is very close to, but not identical with, the result given by Samuelson in [5]. Perhaps it should be noted here that the last statement of the first paragraph of [5] is incorrect.

Uniqueness Under Transformation Criteria, Multivariate Case. For $p > 2$ the following natural criterion is added.

Criterion 5. *Consistency under omission of variables.* Suppose any subset of at least two X_i's is chosen from the original p X_i's. Then the α and λ for this lower dimensional distribution must have as components the corresponding components of α and λ for the original p-variate distribution.

Criterion 5 simply says that the line of organic correlation for any marginal distribution of two or more components must be the projection of the p-dimensional line of organic correlation. Actually, it suffices to state Criterion 5 in terms of pairs of components. We then have

Theorem 2. Under Assumptions 1 and 2 the unique line satisfying Criteria 1 through 5 is the diagonal line of organic correlation.

Proof. By Theorem 1 we have a bivariate diagonal line of organic correlation uniquely determined by each bivariate marginal distribution. There is at most one line in p-dimensional space whose projections on the bivariate coordinate planes will be these uniquely determined lines. By Criterion 5 the p-dimensional diagonal line of organic correlation is in fact this line, and its projections have the proper directions because of the second sentence of Criterion 4.

Without Criterion 5 other λ functions would be possible, for example

$$\lambda_i = [\operatorname{sgn}(\sigma_{1i})]\left[\prod_{j=1}^{p} \frac{|\sigma_{ij}|}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right]\sqrt{\sigma_{ii}}\,. \tag{6}$$

The imposition of further criteria beyond 1 through 5 will in general result in a contradiction, and the nonexistence of any line satisfying all of the criteria. For example, no line satisfies Criteria 1 though 5 and in addition rotates properly under orthogonal transformations of the variables.

If the components of X_i are permuted, then the corresponding components of α and λ for the diagonal line of organic correlation are permuted in the same way. This desirable property follows immediately from the definition of the diagonal line of organic correlation.

The diagonal line of organic correlation depends on the covariances only weakly via their signs. Thus, for example, two p-variate normal distributions having very different ellipsoids of constant density, say one thin and cigar-shaped, the other almost spherical, will give rise to the same line if the means are respectively equal, the signs of the covariances match, and the variances are respectively equal. The descriptive term "diagonal" arises from a geometrical interpretation

mentioned by Jones [4]. For any $c > 0$ consider the ellipsoid in p-space made up of all $x = (x_1, \cdots, x_p)$ satisfying $(x - \mu)\Sigma^{-1}(x - \mu)' = c$. Picture the circumscribed hyper-rectangle about this ellipsoid whose sides are $p - 1$ dimensional hyper-planes perpendicular to the coordinate axes and tangent to the ellipsoid. Then the diagonal line of organic correlation lies along that diagonal of the hyper-rectangle satisfying Criterion 4.

Theorem 1 may be strengthened in two minor ways. First, in Criterion 4 we need not require that no λ_i is zero. Second, the requirement in Assumption 1 that Σ be non-singular may be omitted if it is replaced by (1) the requirement in Assumption 1 that Σ does not degenerate into the zero matrix; (2) the rephrasing of Assumption 2 so that it applies only to covariances, σ_{ij}, for which both σ_{ii} and σ_{jj} are positive; and (3) the restatement of Criterion 4 as follows: $\lambda_i\sigma_{ij}\lambda_j \geq 0$ $(i, j = 1, 2, \cdots, p)$ and if $\sigma_{ii} = 0$ then $\lambda_i = 0$. In order that the definition of the diagonal line of organic correlation given in *Preliminaries* may hold, we suppose the coordinates numbered so that $\sigma_{11} \neq 0$.

Uniqueness under the Prediction Scheme. We now consider the predictive derivation of the diagonal line of organic correlation, using Assumptions 1, 2, plus

Assumption 3. Every bivariate marginal distribution of two components of X is normal.

Let us restrict our attention to lines without vanishing relative trends (i.e., with all $\lambda_i \neq 0$). Later, the effect of permitting zero λ_i's will be discussed briefly.

For simplicity consider first the bivariate case. Suppose that we plan to use the line of organic correlation in the following predictive way. We observe the value of X_1, say x_1, for some organism of the type studied, and we want to predict X_2. First we note that the value of X_2, say x_2, corresponding to x_1 on the line (α, λ) is given by

$$\frac{x_2 - \alpha_2}{\lambda_2} = \frac{x_1 - \alpha_1}{\lambda_1} \tag{7}$$

or

$$x_2 = \alpha_2 + \frac{\lambda_2}{\lambda_1}(x_1 - \alpha_1).$$

Take a positive number β and predict that, for the organism whose X_1 value was observed, its X_2 value lies between $x_2 - \beta\sqrt{\sigma_{22}}$ and $x_2 + \beta\sqrt{\sigma_{22}}$. Repeat this procedure a large number of times and

ask about the probability that the x_2 prediction is correct. This is of course just the probability that

$$| X_2 - \alpha_2 - (\lambda_2/\lambda_1)(X_1 - \alpha_1) | \leq \beta\sqrt{\sigma_{22}} .$$

But $X_2 - \alpha_2 - (\lambda_2/\lambda_1)(X_1 - \alpha_1)$ is normal so that the probability of interest is

$$P_{12}(\beta, \mu, \Sigma, \alpha, \lambda) = \Phi\left\{\frac{\beta\sqrt{\sigma_{22}} - [\mu_2 - \alpha_2 - (\lambda_2/\lambda_1)(\mu_1 - \alpha_1)]}{\sqrt{\sigma_{22} - 2(\lambda_2/\lambda_1)\sigma_{12} + (\lambda_2/\lambda_1)^2\sigma_{11}}}\right\} \tag{8}$$

$$- \Phi\left\{\frac{-\beta\sqrt{\sigma_{22}} - [\mu_2 - \alpha_2 - (\lambda_2/\lambda_1)(\mu_1 - \alpha_1)]}{\sqrt{\sigma_{22} - 2(\lambda_2/\lambda_1)\sigma_{12} + (\lambda_2/\lambda_1)^2\sigma_{11}}}\right\}$$

where Φ is the cumulative distribution function of the unit-normal distribution.

However, we feel that there is no particular reason to suppose that we will always be predicting X_2 on the basis of an observation on X_1 . We might just as well be making a prediction the other way around: that X_1 lies between $x_1 - \beta\sqrt{\sigma_{11}}$ and $x_1 + \beta\sqrt{\sigma_{11}}$, where x_2 is an observation on X_2 and x_1 is obtained from x_2 by equation (7). The probability of interest then is $P_{21}(\beta, \mu, \Sigma, \alpha, \lambda)$ defined by (8) with subscripts "1" and "2" interchanged.

Suppose that with equal likelihood we will be asked to predict X_2 on the basis of an observation of X_1 and vice versa, and that the predictions must be based upon a single line in the manner described above. Then one might reasonably wish to choose this line so that the average probability of a correct prediction:

$$(1/2)\{P_{12}(\beta, \mu, \Sigma, \alpha, \lambda) + P_{21}(\beta, \mu, \Sigma, \alpha, \lambda)\} \tag{9}$$

is a maximum.

A familiar property of the unit-normal distribution is that the probability of an interval of fixed length is maximized when its midpoint is zero. Hence, whatever be λ_1 and λ_2 , (9) is maximized with respect to α_1 , α_2 when

$$\lambda_1(\mu_2 - \alpha_2) - \lambda_2(\mu_1 - \alpha_1) = 0. \tag{10}$$

Hence we may write P_{ij} as

$$P_{ii}(\beta, \Sigma, \lambda) = 2\Phi\left\{\frac{\beta\sqrt{\sigma_{ii}}}{\sqrt{\sigma_{ii} - 2(\lambda_i/\lambda_i)\sigma_{ii} + (\lambda_i/\lambda_i)^2\sigma_{ii}}}\right\} - 1 \tag{11}$$

which depends neither on μ nor α. The maximizing λ_1 , λ_2 do not depend on μ and by (10) the maximizing α_i's are of form $\mu_i + t\lambda_i$.

The final part of this discussion is the demonstration that $P_{12}(\beta, \Sigma, \lambda) + P_{21}(\beta, \Sigma, \lambda)$ has a single maximum taken on if and only if $\lambda_2/\lambda_1 = \text{sgn}(\sigma_{12})\sqrt{\sigma_{22}}/\sqrt{\sigma_{11}}$. In other words, the probability of a correct prediction is maximized if the predicting line is the diagonal line of organic correlation. If we set $r = (\lambda_1/\lambda_2)(\sqrt{\sigma_{22}}/\sqrt{\sigma_{11}})$, the quantity to be maximized with respect to r is

$$(12) \qquad \Phi\left\{\frac{\beta}{\sqrt{1 - 2r\rho + r^2}}\right\} + \Phi\left\{\frac{\beta}{\sqrt{1 - 2r^{-1}\rho + r^{-2}}}\right\} - 1$$

where $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, the correlation coefficient between X_1 and X_2. As r approaches $\pm\infty$ or 0, (12) approaches $\Phi(\beta) - \frac{1}{2}$. Note that given $|r|$ that sign for r equal to the sign of σ_{12} gives the greater value to (12). Hence we may as well set $R = |r|$ and maximize

$$(13) \qquad \Phi\left\{\frac{\beta}{\sqrt{1 - 2|\rho|R + R^2}}\right\} + \Phi\left\{\frac{\beta}{\sqrt{1 - 2|\rho|R^{-1} + R^{-2}}}\right\} - 1.$$

It is a straightforward exercise to show, by examination of the sign of the derivative of (13), that the maximum value of (13) is attained just at $R = 1$. This means that the maximum of (12) is attained when and only when $\lambda_2/\lambda_1 = \text{sgn}(\sigma_{12})\sqrt{\sigma_{22}}/\sqrt{\sigma_{11}}$. The maximized average probability of a correct prediction is just

$$(14) \qquad 2\Phi\left\{\frac{\beta}{\sqrt{2(1 - |\rho|)}}\right\} - 1$$

The assumption of non-singularity may be dropped providing that both σ_{ii} are > 0. The proof goes through as before and of course when $|\rho| = 1$ the value of (14) turns out to be unity. When one, but not both, variances are zero and we permit the corresponding λ_i to be zero, the result still holds.

If in general we attempt to maximize over all possible λ's, not just those with the property: $\sigma_{ii} > 0$ implies $\lambda_i \neq 0$, the situation becomes more complex. In the first place, it is necessary to state precisely how predictions are then to be made. Secondly, having made this precise, in cases of small correlation, there may not be a unique maximizing line, or if there is it may have one λ_i zero even through Σ is non-singular. I believe that this is really of little interest, however, since the biologist is usually not concerned with lines having a zero λ_i.

By considering the marginal bivariate distributions separately, the above manipulations carry over to the multivariate case. We may let the numbers determining interval length be different for each unordered pair (X_i, X_j), and we maximize the probability of a correct

prediction averaged over all possible choices of ordered pairs of components (X_i , X_j). This gives us

Theorem 3. Suppose that Assumptions 1, 2, and 3 hold. Suppose further that we choose numbers $\beta_{ij} = \beta_{ji}$ $(i, j = 1, \cdots , p; i \neq j)$ and that for any line (α,λ) we consider the probability that

$$| (X_J - \alpha_J) - (\lambda_J/\lambda_I)(X_I - \alpha_I) | \leq \beta_{IJ}\sqrt{\sigma_{JJ}}$$

where the ordered subscript couple (I,J) is chosen from the $p(p-1)$ possibilities at random, each with equal probability.[4]

Then the line which maximizes this probability under the restriction $\lambda_i \neq 0$ $(i = 1, \cdots , p)$ is the diagonal line of organic correlation. Further, the maximum probability resulting from the use of this line is

$$\frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j=1}^{p} \left[2\Phi\left\{ \frac{\beta_{ij}}{\sqrt{2(1 - | \rho_{ij} |)}} \right\} - 1 \right]$$

where $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$.

If $\beta_{ij} \neq \beta_{ji}$, or if predictions of X_j from X_i are more likely than from X_i to X_j , then the maximizing line changes and is in general difficult to describe explicitly. One simple and standard case is when $p = 2$ and we always predict X_2 from X_1 . Then the best line is the ordinary regression of X_2 on X_1 .

Minimizing the Expected Value of a Right Triangle's Area. Teissier [3] and others before him suggest the same (α,λ), but with a different and more ad hoc motivation than the two presented thus far. They consider the marginal bivariate distribution of, say, (X_1 , X_2) and measure the deviation of an observation (x_1 , x_2) from the projection of (α,λ) on the (x_1 , x_2) plane by the area of the right triangle which has (x_1 , x_2) as a vertex and has sides, parallel to the axes, running from this vertex to the line (α,λ). They suggest that that line (α,λ) minimizing the expected value of this area seems optimal. The requirement, then, is that

$$E\left\{ \left| \left[(X_2 - \alpha_2) - \frac{\lambda_2}{\lambda_1} (X_1 - \alpha_1) \right] \left[(X_1 - \alpha_1) - \frac{\lambda_1}{\lambda_2} (X_2 - \alpha_2) \right] \right| \right\}$$

be minimized. If we restrict ourselves to non-zero λ_1 and λ_2 , this comes to minimizing

$$E\left\{ \left[\sqrt{\left| \frac{\lambda_2}{\lambda_1} \right|} (X_1 - \alpha_1) - \operatorname{sgn}(\sigma_{12}) \sqrt{\left| \frac{\lambda_1}{\lambda_2} \right|} (X_2 - \alpha_2) \right] \right\}^2 .$$

[4]Actually it would suffice for the (I,J)'s to be chosen so that it is just as probable that (I,J) be (i,j) as that it be (j,i).

with respect to α_1 , α_2 , λ_1 , λ_2 . It is easily shown that the minimum is attained for

$$\alpha_1 = \mu_1 , \qquad \alpha_2 = \mu_2 , \qquad \lambda_1/\lambda_2 = \operatorname{sgn}(\sigma_{12})\sqrt{\sigma_{11}/\sigma_{22}} .$$

Extending this minimization to all the pairs (X_i , X_j), we may state the following formal result

Theorem 4. (Teissier and others) Under Assumptions 1 and 2 the unique line which simultaneously minimizes the $p(p-1)$ quantities

$$E\left\{\left|\left[(X_j - \alpha_j) - \frac{\lambda_j}{\lambda_i}(X_i - \alpha_i)\right]\left[(X_i - \alpha_i) - \frac{\lambda_i}{\lambda_j}(X_j - \alpha_j)\right]\right|\right\}$$

under the restrictions $\lambda_i \neq 0$ $(i = 1, \cdots , p)$, is the diagonal line of organic correlation.

The restriction on the λ_i's is only for convenience of notation and statement, since for any bivariate marginal distribution the expected triangle area approaches infinity as either of the two associated λ_i's approaches zero while the other is held fixed.

The Diagonal Line of Organic Correlation in Terms of the Quantiles of the Marginal Distributions. Greenall [7] gives an interesting characterization for the diagonal line of organic correlation along the following lines. Suppose that all the correlations are positive. Impose Criterion 4 and hence consider all lines (α,λ) having their λ_i's positive. Now set up the following requirement:

$$\Pr\{X_1 \leq \alpha_1 + \lambda_1 t\} = \Pr\{X_2 \leq \alpha_2 + \lambda_2 t\} = \cdots$$
$$= \Pr\{X_p \leq \alpha_p + \lambda_p t\} \qquad (\text{all } t)$$

where "Pr $\cdots$ " means the probability of $\cdots$.

This requirement states that we seek a line that levels out, so to speak, the quantiles of the marginal distributions; that is a line such that at any point the marginal cumulative distributions, component-wise, are all equal. In other words the requirement is that the random variables $(X_i - \alpha_i)/\lambda_i$ all have the same distribution. A case of particular interest is that in which the marginal distributions of the X_i's belong to a location-scale family of distributions having means and variances, that is in which there is a random variable Y, having zero mean and unit variance, such that the distribution of X_i is that of $\sqrt{\sigma_{ii}}\, Y + \mu_i$. If no σ_{ii} is zero the requirement then is that the random variables $(\sqrt{\sigma_{ii}}/\lambda_i)Y + (\mu_i - \alpha_i)/\lambda_i$ all have the same distribution. A glance at the first two moments of these random variables shows that this can happen if and only if the α_i's and λ_i's are those of the diagonal line

of organic correlation. Note that the case of marginal normality is covered by the above remarks, but so are many other cases.

If the correlations are not all positive we may still carry out the above argument by transforming to the positive correlation case via appropriate -1 multiplications, using the above argument, then transforming back under the requirement of Criterion 3.

This discussion is summarized by

Theorem 5. (Greenall). Assume that the X_i's all belong to a location-scale family and possess means and positive variances. Impose Assumption 2, Criterion 3,[5] and Criterion 4. Then the diagonal line of organic correlation is the unique line having the property

$$\Pr\{X_1 \leq \alpha_1 + \lambda_1 t\} = \Pr\{X_2 \leq \alpha_2 + \lambda_2 \operatorname{sgn}(\sigma_{12})t\}$$
$$\cdots = \Pr\{X_p \leq \alpha_p + \lambda_p \operatorname{sgn}(\sigma_{1p})t\}$$

Greenall's paper contains still another characterization of the diagonal line of organic correlation. In addition it argues for the utility of a single representing line in certain psychological contexts.

Acknowledgements. The questions considered here were suggested to me by Professors E. C. Olson and R. L. Miller during discussions of applications of statistics to paleontology. I should like to thank Professors Olson and Miller for their helpful comments made after reading a draft of this paper. Further, I should like to thank Professor T. Koopmans for references to the relevant econometric literature, and the referee for valuable suggestions both as to content and presentation.

REFERENCES

[1] K. A. Kermack and J. B. S. Haldane, "Organic Correlation and Allometry," *Biometrika* (1950), 37, 30.

[2] T. C. Koopmans and O. Reiersøl, "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics* (1950), 21, No. 2, 165.

[3] G. Teissier, "La Relation d'Allometrie: Sa Signification Statistique et Biologique," *Biometrics* (1948), 4, No. 1, 14.

[4] H. F. Jones, "Some Geometrical Considerations in the General Theory of Fitting Lines and Planes," *Metron*, 13, No. 1, 21.

[5] P. A. Samuelson, "A Note on Alternative Regressions," *Econometrica* (1942), 10, No. 1, 80.

[6] R. Frisch, "Statistical Confluence Analysis by Means of Complete Regression Systems," Universitets Økonomiske Institutt, Publikasjon Nr. 5, Oslo, 1934.

[7] P. D. Greenall, "The Concept of Equivalent Scores in Similar Tests," *British Journal of Psychology, Statistical Section*, (1949), 2, 30.

[5] Actually the b_i's in Criterion 3 here need take only the values ± 1.

VARIANCE OF A WEIGHTED MEAN*†

PAUL MEIER

School of Hygiene and Public Health
The Johns Hopkins University

1. *Introduction*

In many areas of statistical practice the problem arises of combining several estimates of an unknown quantity to obtain an estimate of improved precision.

For example, suppose two or more technicians have performed assays on several samples from a homogeneous material. It is desired to average their results in the best manner possible, allowing for the fact that technicians differ in precision. In general the relative precisions are not known exactly, but estimates are available from current or previous experimental data.

A similar problem arises in the analysis of incomplete block experiments. The "intra-block" and "inter-block" estimates of varietal means have different variances, and the recovery of "inter-block information" is an attempt to combine these estimates in the most efficient manner. Although similar methods are applicable, the experimental design problem differs from the simple case of weighted means in several important respects. An investigation of this problem in the case of simple lattice designs will be presented in a subsequent paper.

The model for the first problem may be described as follows. Let $u_1 , \cdots , u_k$ be k estimates of a parameter μ, being independently and normally distributed with mean μ and variances $\sigma_1^2 , \cdots , \sigma_k^2$. Also let $s_1^2 , \cdots , s_k^2$ be independent unbiased estimates of the σ_i^2 having

*This paper is a revision of Part I of a thesis submitted to Princeton University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

†Paper No. 287 from the Department of Biostatistics.

mean square distributions on $n_1, \cdots, n_k$ degrees of freedom respectively. We define notation as follows:

$$w_i = \frac{1}{\sigma_i^2}, \qquad w = \sum w_i, \qquad \theta_i = \frac{w_i}{w}.$$

Now if the σ_i^2 are given, the best estimate of μ is known to be $\sum \theta_i u_i$ and the variance of this estimate is $1/w$. For the case in which the variances are estimated we consider the analogous estimate, $\bar{\mu} = \sum \hat{\theta}_i u_i$, where a "$\hat{}$" denotes the replacement of σ_i^2 by s_i^2 in the formula for the indicated quantity. In particular we investigate $V(\bar{\mu})$, the variance of $\bar{\mu}$. It should be noted that the $\hat{\theta}_i$ are not the maximum likelihood weights. However, these weights provide an asymptotically efficient estimate of μ when the n_i are large, and have the advantage of simplicity of calculation, whereas the maximum likelihood weights require iterative procedures for their determination.

Our investigation gives first order asymptotic results for a fixed number, k, of estimates, with errors of order $\sum 1/n_i^2$. The basic results are

(1) $$\text{Var}\{\bar{\mu}\} = \frac{1}{w}\left\{1 + 2\sum_{i=1}^{k} \frac{1}{n_i}\theta_i(1 - \theta_i) + 0\left(\sum_{i=1}^{k} \frac{1}{n_i^2}\right)\right\}$$

(2) An approximately unbiased estimate of $\text{Var}\{\bar{\mu}\}$ is given by

$$V^* = \frac{1}{\hat{w}}\left\{1 + 4\sum_{i=1}^{k} \frac{1}{n_i}\hat{\theta}_i(1 - \hat{\theta}_i)\right\}$$

(3) V^* is distributed approximately as a mean square with f degrees of freedom where

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{\theta_i^2}{n_i}$$

The reader whose major interest is in applications should refer to the numerical example, Section 3, where the procedural details are described.

2. *Relation to Previous Investigations*

The problem of weighted means was investigated by Cochran in 1937 [5] for the case of a large number of estimates and constant $n_i = n$. It was shown that in this case $\text{Var}\{\bar{\mu}\} \approx 1/w\{1 + 2/(n - 4)\}$ and that an approximately unbiased estimate of this variance is given by $1/\hat{w}\{1 + 4/(n - 4)\}$. More recently numerical investigations of this case for different values of k have been carried out by Sarah Porter

Carroll [3] and by Carroll and Cochran [4]. They also give an empirical formula for estimating $\mathrm{Var}\{\bar{\mu}\}$ when k is small.

Another investigation was made by Neyman and Scott [8]. They studied the maximum likelihood estimate of μ for the case of large k, but relaxed the requirement of equality of the n_i . It was shown that the maximum likelihood estimator is not efficient. A similar but more efficient estimator was exhibited. Estimates of the variance were not considered.

In addition it should be noted that our methods are similar to those used by Welch [11] to investigate the Behrens-Fisher problem.

3. *Numerical Example*

In an example given by Snedecor [10] the data from four experiments are used to estimate the percentage of albumin in the plasma protein of normal human subjects. The data are shown in table I.

TABLE I

Experimenter	Number of subjects (m_i)	Degrees of freedom (n_i)	Mean percentage (u_i)	Population variance estimate
A	12	11	62.3	12.986
B	15	14	60.3	7.840
C	7	6	59.5	33.433
D	16	15	61.5	18.513

If it be assumed that all four estimates have a common mean[1] and the same population variance (Bartlett's test for homogeniety of variances gives a value of $\chi^2 = 5.14$ on 3 degrees of freedom) the appropriate estimate of the mean obtained by pooling the data is 61.05%, with an estimated variance of 0.3178 on 46 degrees of freedom. However, it is evident that the results are consistent with a considerable divergence in the true population variances. If we assume only that the four populations have a common mean, but possibly different variances, we may proceed as follows.[2]

[1]It has been pointed out to me independently by Professors C. I. Bliss and Margaret Merrell that this assumption is likely to be invalid in experiments of this type. Experimenters with large variances may tend also to have large biases. The problem of weighted means when biases are present is not considered in this paper. This problem is discussed by Cochran [5], and by Yates and Cochran [12].

[2]The accuracy of the correction term has not been determined for the general case. Since one of the estimates is based on only 6 degrees of freedom the example should be viewed only as an illustration of the method.

(a) Find the estimated variances of the means by dividing the estimated population variances by the sample sizes,

$$s_i^2 = \frac{\text{pop'n. var. estimate}}{m_i}$$

(b) Find the weights by inverting these variance estimates,

$$w_i = \frac{1}{s_i^2}, \qquad w \text{ is the sum of the } w_i .$$

Steps (a) and (b) may be tabulated as shown in Table II.

TABLE II

Experimenter	Number of subjects (m_i)	Variance estimates for means (s_i^2)	Weights (w_i)
A	12	1.0822	0.9241
B	15	0.5227	1.9133
C	7	4.7761	0.2094
D	16	1.1571	0.8643
			$w = 3.9110$

(c) Calculate the weighted mean

$$\bar{\mu} = \frac{\sum w_i u_i}{w}$$

$$= \frac{1}{3.9110}[(0.9241)(62.3) + (1.9133)(60.3) + (0.2094)(59.5) + (0.8643)(61.5)]$$

$$= 60.99$$

(d) Calculate the estimated variance of $\bar{\mu}$ from (2), which may be written

$$V^* = \frac{1}{w}\left[1 + \frac{4}{w^2}\sum \frac{1}{n_i} w_i(w - w_i)\right]$$

$$= 0.2557\left\{1 + \frac{4}{(3.9110)^2}\left[\frac{1}{11}(0.9241)(2.9869) + \frac{1}{14}(1.9133)(1.9977) + \frac{1}{6}(0.2094)(3.7016) + \frac{1}{15}(0.8643)(3.0467)\right]\right\}$$

$$= 0.3111$$

(e) Calculate the estimated equivalent degrees of freedom from (3), which may be written

$$\hat{f} = \frac{\hat{w}^2}{\sum \frac{\hat{w}_i^2}{n_i}} = \frac{(3.9110)^2}{\frac{1}{11}(0.9241)^2 + \frac{1}{14}(1.9133)^2 + \frac{1}{6}(0.2094)^2 + \frac{1}{15}(0.8643)^2}$$

$$= 38.6$$

(Note that in (d) and (e) the sample degrees of freedom rather than the sample sizes are used.)

Thus, finally, $\bar{\mu} = 60.99$ and $V^* = 0.3111$ on 38.6 degrees of freedom.

We may now compare three possible methods of treating this problem, assuming no biases are present.

(1) Assume all population variances are equal, i.e. the four samples come from a single population (equal variance method). This is the treatment used by Snedecor.

(2) Allow for the possibility of different population variances, and use the four variance estimates as if they were exactly equal to the true variances (unequal variance, uncorrected method).

(3) Allow for the possibility of different population variances, but use the above corrections to allow for the sampling variability of the weights (unequal variance, corrected method).

The results of the three methods may be summarized as follows.

TABLE III

Method	Estimate of μ	Variance of estimate	Degrees of freedom
(1) equal variance	61.05	0.3178	46
(2) uncorrected	60.99	0.2557	—
(3) corrected	60.99	0.3111	38.6

We see that the estimates of μ differ only trivially; the estimate for the equal variance method is 0.06 greater than the estimate for methods (2) and (3), a small fraction of the estimated standard deviation.

However, the variance estimates differ substantially. Methods (1) and (3) agree fairly closely, differing by approximately 2%, but the uncorrected estimate is 0.0554 less than the corrected version, a deficiency of 18%.

With respect to the stability of the variance estimate the uncorrected method yields no measure, although the 46 DF for method (1) might be taken for an upper bound. The drop from 46 DF to 38.6 DF indicated by method (3) has a negligible effect on the 5% and 1% levels of Student's t distribution.

The above example illustrates the fact that in the analysis for the case of unequal variances the required correction term may be substantial. The near equality of the results of methods (1) and (3) may not be surprising in view of the fact that Bartlett's test provided no evidence against the hypothesis of equal variances (see p. 61). Method (3), however, is the more conservative procedure and is applicable whenever the assumption of equal variances is doubtful, whether or not Bartlett's test gives a significant result.

The remainder of this paper is devoted to proving the basic relations stated in the introduction (section 4) and to an examination of the special case of two samples (section 5). The results are summarized in section 6.

4. *The General Case*

We are concerned with the variance of $\bar{\mu} = \sum_{i=1}^{k} \hat{\theta}_i u_i$, where the u_i are independently and normally distributed with mean μ and variances σ_i^2.

$$\hat{\theta}_i = \frac{(s_i^2)^{-1}}{\sum_{i=1}^{k} (s_i^2)^{-1}}$$

where $s_1^2, \cdots, s_k^2$ are unbiased estimates of $\sigma_1^2, \cdots, \sigma_k^2$ independently distributed as mean squares on $n_1, \cdots, n_k$ degrees of freedom. We shall be interested also in the distribution of estimates of Var $\{\bar{\mu}\}$.

The main tool of our investigation is the following theorem, an analogue of which is used by Welch [11].

Theorem If $x_1, \cdots, x_k$ are independently distributed with density functions

$$f_{n_i}(x_i) = \frac{\left(\frac{n_i}{2}\right)^{n_i/2}}{\Gamma\left(\frac{n_i}{2}\right)} x_i^{(n_i/2)-1} e^{-n_i x_i/2} \quad (0 \leq x_i < \infty)$$

and $R(x_1, \cdots, x_k)$ is a rational function with no singularities for $0 < x_1, \cdots, x_k < \infty$, then Ave $\{R(x_1, \cdots, x_k)\}$ can be expanded in an asymptotic series in the $1/n_i$. In particular

$$\text{Ave } \{R(x_1, \cdots, x_k)\}$$

$$= R(1, \cdots, 1) + \sum_{i=1}^{k} \frac{1}{n_i} \frac{\partial^2 R}{\partial x_i^2}\Big|_{(1,\cdots,1)} + 0\left(\sum \frac{1}{n_i^2}\right).$$

The theorem is proved by means of the method of steepest descents. The function $R(x_1, \cdots, x_k)$ need not, in fact, be rational. If it has a Taylor series valid in the neighborhood of $(1, \cdots, 1)$ and does not go to infinity too rapidly the theorem remains valid. All the functions which we shall consider are rational.

a. *Variance of* $\bar{\mu}$

It will now be convenient to define quantities x_i by $s_i^2 = \sigma_i^2 x_i$. The x_i are then distributed with the density functions $f_{n_i}(x_i)$. We then have

$$\bar{\mu} = \sum \hat{\theta}_i u_i = \frac{\sum w_i \dfrac{u_i}{x_i}}{\sum \dfrac{w_i}{x_i}}$$

Since the u_i and x_i are independent we may take average values successively—first with respect to the u_i holding the x_i fixed, and then with respect to the x_i. Therefore we may write the conditional variance of $\bar{\mu}$ given the x_i as

$$V(\bar{\mu} \mid x_i) = \frac{\sum \dfrac{w_i}{x_i^2}}{\left(\sum \dfrac{w_i}{x_i}\right)^2}$$

Now $V(\bar{\mu}|x_i)$ clearly satisfies the conditions of the theorem, so we can write the variance of $\bar{\mu}$ as

$$V(\bar{\mu}) = V(\bar{\mu} \mid 1, \cdots, 1) + \sum \frac{1}{n_i} \frac{\partial^2 V(\bar{\mu} \mid x_i)}{\partial x_i^2}\Big|_{(1,\cdots,1)} + 0\left(\sum \frac{1}{n_i^2}\right)$$

$$= \frac{1}{w}\left\{1 + 2 \sum \frac{1}{n_i} \theta_i(1 - \theta_i) + 0\left(\sum \frac{1}{n_i^2}\right)\right\}$$

b. *Estimation of* $V(\bar{\mu})$

A natural estimate of $V(\bar{\mu})$ would be $1/\hat{w}\,\{1 + 2\sum (1/n_i)\hat{\theta}_i(1 - \hat{\theta}_i)\}$. However, this estimate has, asymptotically, a negative bias which is approximately equal in magnitude to the correction term. This may

be seen as follows. Application of the theorem to the first term of the above estimate, namely $1/\hat{w}$, yields.

$$\text{Ave}\left\{\frac{1}{\hat{w}}\right\} = \frac{1}{w}\left\{1 - 2\sum \frac{1}{n_i}\theta_i(1-\theta_i) + 0\left(\sum \frac{1}{n_i^2}\right)\right\}$$

Obviously to obtain an estimate with bias of order $0(\sum 1/n_i^2)$ we must double the correction term. Hence

$$V^* = \frac{1}{\hat{w}}\left\{1 + 4\sum \frac{1}{n_i}\hat{\theta}_i(1-\hat{\theta}_i)\right\}$$

is an estimate of $V(\bar{\mu})$ with bias of order $0(\sum 1/n_i^2)$.

c. *Stability of the estimate of* $V(\bar{\mu})$

One measure of the stability of V^*, our estimate of $V(\bar{\mu})$, is its variance. To first order terms in the $1/n_i$ it is clear that the variance of V^* is the same as that of $1/\hat{w}$. Since

$$\text{Ave}\left\{\left(\frac{1}{\hat{w}}\right)^2\right\} = \left(\frac{1}{w}\right)^2\left\{1 - 2\sum \frac{1}{n_i}\theta_i(2-3\theta_i) + 0\left(\sum \frac{1}{n_i^2}\right)\right\} \quad \text{and}$$

$$\text{Ave}^2\left(\frac{1}{\hat{w}}\right) = \left(\frac{1}{w}\right)^2\left\{1 - 2\sum \frac{2}{n_i}\theta_i(1-\theta_i) + 0\left(\sum \frac{1}{n_i^2}\right)\right\}$$

we may write

$$\text{Var}\{V^*\} = 2\left(\frac{1}{w}\right)^2\left\{\sum \frac{\theta_i^2}{n_i} + 0\left(\sum \frac{1}{n_i^2}\right)\right\}$$

By analogy with a mean square distribution we are tempted to use the designation of *equivalent degrees of freedom* for the quantity

$$f = \frac{1}{\sum \frac{\theta_i^2}{n_i}}.$$

This quantity is the same as that which appears in the study of variance components [9]. To the order of accuracy considered here $\hat{f}$ is a suitable estimate of f.

The obvious question which now arises is whether the distribution of V^* based on small n_i is satisfactorily approximated by a mean square distribution with f degrees of freedom. More particularly, will the tabulated percent points of t-tests based on V^* with $\hat{f}$ degrees of freedom be close to the true percent points? These questions cannot be answered satisfactorily in the absence of detailed investigation of the exact distributions. Pending such investigation we recommend treating

$\bar{\mu}/\sqrt{V^*}$ as a t variate with $\hat{f}$ degrees of freedom in preference to using the uncorrected quantity $\bar{\mu}/\sqrt{1/\hat{w}}$.

5. *Case of two means* ($k = 2$)

When only two estimates are to be combined the calculation of variances is simpler, permitting an exact evaluation of $V(\bar{\mu})$ and an investigation of the bias of V^*. The formula for $V(\bar{\mu} \mid x_i)$ reduces to

$$V(\bar{\mu} \mid x_i) = \frac{\frac{w_1}{x_1^2} + \frac{w_2}{x_2^2}}{\left(\frac{w_1}{x_1} + \frac{w_2}{x_2}\right)^2} = \frac{1}{w}\left\{1 + w_1 w_2 \frac{(x_1 - x_2)^2}{(w_1 x_2 + w_2 x_1)^2}\right\}$$

This is a function of the ratio x_1/x_2 which has the well known F distribution, enabling us to write

$$V(\bar{\mu}) = \text{Ave}\,\{V(\bar{\mu} \mid x_i)\}$$

$$= \frac{1}{w}\left\{1 + \frac{w_2/w_1}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^\infty \frac{\left(1 - \frac{n_2}{n_1} t\right)^2}{(1 + \gamma t)^2} \frac{t^{(n_1/2)-1}}{(1+t)^{(n_1+n_2)/2}}\, dt\right\}$$

where

$$\gamma = \frac{n_2 w_2}{n_1 w_1}$$

We refer to the second term within the braces as *the fractional correction to* $1/w$, i.e. the fraction by which $1/w$ must be increased to yield $V(\bar{\mu})$. This fractional correction to $1/w$ will be denoted by the letter C.

a. *Bounds on the increase in variance*

The variance of $\bar{\mu}$ is most conveniently described in terms of C, the fractional correction to $1/w$ which is defined above. We lose no generality if we assume that the notation has been assigned in such a manner that $\gamma \geq 1$. Using the relations

$$\frac{1}{\gamma^2(1+t)^2} \leq \frac{1}{(1+\gamma t)^2} \leq \frac{1}{(1+t)^2}$$

we find that

$$\frac{2}{\gamma(n_1 + n_2 + 2)} \leq C \leq \frac{2\gamma}{n_1 + n_2 + 2}$$

with equality when $\gamma = 1$. These bounds are close only when γ is in

the neighborhood of one. [Note: An upper bound to $V(\bar{\mu})$ when $k > 2$ is given by

$$V(\bar{\mu}) < \frac{1}{w}\left\{1 + \sum_{i<j} C_{ij}\right\} \quad \text{where} \quad C_{ij} = \frac{2\gamma_{ij}}{n_i + n_j + 2}$$

and

$$\gamma_{ij} = \max\left(\frac{n_i w_i}{n_j w_j}, \frac{n_j w_j}{n_i w_i}\right)\Bigg].$$

b. *Case with degrees of freedom proportional to variances*

A partial check on the accuracy of our approximation can be made by comparing the results with the exact values in the case $n_1/n_2 = \sigma_1^2/\sigma_2^2$, or equivalently, $n_1 w_1 = n_2 w_2$. In this case $\gamma = n_2 w_2 / n_1 w_1 = 1$ and the above relation becomes an equality. Thus, for this case

$$V(\bar{\mu}) = \frac{1}{w}\left\{1 + \frac{2}{n_1 + n_2 + 2}\right\}$$

Our approximation is

$$V(\bar{\mu}) \approx \frac{1}{w}\left\{1 + 2\sum \frac{1}{n_i}\theta_i(1 - \theta_i)\right\} = \frac{1}{w}\left\{1 + \frac{2}{n_1 + n_2}\right\}$$

The approximation is too large by the amount

$$\frac{4}{(n_1 + n_2)(n_1 + n_2 + 2)}$$

Of more direct interest is the bias in the observed quantity

$$V^* = \frac{1}{\hat{w}}\left\{1 + 4\sum \frac{1}{n_i}\hat{\theta}_i(1 - \hat{\theta}_i)\right\} = \frac{1}{\hat{w}}\left\{1 + 4\hat{\theta}_1\hat{\theta}_2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}$$

Ave $\{V^*\}$ can be evaluated straightforwardly, giving the result

$$\text{Ave}\,\{V^*\} = \frac{1}{w}\left\{1 + \frac{2}{n_1 + n_2 + 2} + 8\,\frac{(n_1 + n_2)^3 - 5n_1n_2(n_1 + n_2) + 2(n_1 + n_2)^2 - 12n_1n_2}{n_1n_2(n_1 + n_2 + 2)(n_1 + n_2 + 4)(n_1 + n_2 + 6)}\right\}$$

For $n_1 = n_2 = n$ this reduces to

$$\text{Ave}\,\{V^*\} = \frac{1}{w}\left\{1 + \frac{1}{n + 1} - \frac{2}{(n + 1)(n + 3)}\right\}$$

Thus the bias is seen to be rather small, being already less than 5% when $n_1 = n_2 = 4$.

c. *Case with large variance ratio*

For $\gamma > 1$ the quantity $V(\bar{\mu} \mid x_i)$ can be expanded in a power series

in $1/\gamma$. Term by term integration to evaluate $V(\bar{\mu})$ gives an asymptotic series valid for large γ. Thus

$$C \sim \frac{2}{\gamma}\frac{n_1^2 + n_1n_2 + 4n_2}{n_1(n_1 - 2)(n_1 - 4)} - \frac{4}{\gamma^2}\frac{n_1^2n_2 + n_1n_2^2 + 4n_1^2 + 12n_2^2 + 12n_1n_2}{n_1(n_1 - 2)(n_1 - 4)(n_1 - 6)} + \cdots,$$

the error being less than the first term ignored. Unfortunately for moderate size n_1 , n_2 the series is useful only for rather large γ (e.g. for $n_1 = n_2 = 10$ we require $\gamma \approx 15$ to make the error less than 0.05).

The average value of V^* can be obtained similarly. The expressions are rather complex and uninteresting so we omit them. However, it appears that for large variance ratio V^* slightly underestimates $V(\bar{\mu})$.

d. *Exact calculation of* $V(\bar{\mu})$ *for* $n_1 = n_2 = 1, 2, 3, 4, 5, 6$.

For the sake of simplicity the detailed investigation of $V(\bar{\mu})$ will be restricted to the case $n_1 = n_2 = n$. (This restriction is not essential. A similar analysis can be made when $n_1 \neq n_2$). In this case the fractional correction to $1/w$ is

$$C_n(\alpha) = \frac{\alpha}{\beta\left(\frac{n}{2}, \frac{n}{2}\right)}\int_0^\infty \frac{(1-t)^2}{(1+\alpha t)^2}\frac{t^{(n/2)-1}}{(1+t)^n}\,dt, \quad \text{where } \alpha = \frac{w_2}{w_1} = \frac{\sigma_1^2}{\sigma_2^2}$$

Now it is evident that if σ_1^2 and σ_2^2 are interchanged the fractional correction to $1/w$ will remain the same. Thus $C_n(1/\alpha) = C_n(\alpha)$, a fact which can be verified by direct substitution. It is sufficient, therefore, to calculate $C_n(\alpha)$ for $\alpha \geq 1$.

An indication of the behavior of $C_n(\alpha)$ in the neighborhood of $\alpha = 1$ is given by its derivatives. We have

$$C_n\Big]_{\alpha=1} = \frac{1}{n+1}$$

$$\frac{dC_n}{d\alpha}\Big]_{\alpha=1} = 0$$

$$\frac{d^2C_n}{d\alpha^2}\Big]_{\alpha=1} = -\frac{(n-6)}{2(n+1)(n+3)}$$

It appears that at $\alpha = 1$, $C_n(\alpha)$ has a local minimum when $n < 6$ and a maximum when $n > 6$. For $n = 6$

$$\frac{d^2C_6}{d\alpha^2}\Big]_{\alpha=1} = \frac{d^3C_6}{d\alpha^3}\Big]_{\alpha=1} = 0, \qquad \text{but} \qquad \frac{d^4C_6}{d\alpha^4}\Big]_{\alpha=1} = -\frac{2}{1155}$$

so $C_6(\alpha)$ also has a maximum at $\alpha = 1$.

Now to evaluate $C_n(\alpha)$ for small n we note that the integrand is a rational function of $\sqrt{t}$ and can be integrated by the methods of elementary calculus. This procedure becomes rather tedius for even moderate n and the form of the result, being rather complex, does not seem to warrant a considerable listing. The formulas and numerical results for $n = 1, 2, 3, 4, 5, 6$ are shown in Tables IV and V

TABLE IV

FRACTIONAL CORRECTION TO $1/w$

n	$C_n(\alpha)$	$\lim_{\alpha \to 1}$	Asymptotic value for large α
1	$\dfrac{\sqrt{\alpha}\,(\alpha+1)(\alpha^2-6\alpha+1)+8\alpha^2}{2\alpha(\alpha-1)^2}$	$\frac{1}{2}$	$\frac{1}{2}\sqrt{\alpha}$
2	$\dfrac{(\alpha^2+6\alpha+1)}{(\alpha-1)^2}-\dfrac{4\alpha(\alpha+1)}{(\alpha-1)^3}\ln\alpha$	$\frac{1}{3}$	1
3	$\dfrac{4\sqrt{\alpha}\,(\alpha+1)(\alpha^2+10\alpha+1)}{(\alpha-1)^4}-\dfrac{4\alpha(5\alpha^2+14\alpha+5)}{(\alpha-1)^4}$	$\frac{1}{4}$	$\dfrac{4}{\sqrt{\alpha}}$
4	$\dfrac{6\alpha(\alpha+1)(\alpha^2+6\alpha+1)}{(\alpha-1)^5}\ln\alpha-\dfrac{4\alpha(5\alpha^2+14\alpha+5)}{(\alpha-1)^4}$	$\frac{1}{5}$	$6\dfrac{\ln\alpha}{\alpha}$
5	$\dfrac{4}{3}\dfrac{\alpha}{(\alpha-1)^6}(7\alpha^4+148\alpha^3+330\alpha^2+148\alpha+7)$ $-\dfrac{64}{3}\dfrac{\sqrt{\alpha}}{(\alpha-1)^6}\alpha(\alpha+1)(3\alpha^2+14\alpha+3)$	$\frac{1}{6}$	$\dfrac{28}{3\alpha}$
6	$4\dfrac{\alpha}{(\alpha-1)^6}(\alpha^4+41\alpha^3+96\alpha^2+41\alpha+1)$ $-60\dfrac{\alpha^2}{(\alpha-1)^7}(\alpha+1)(\alpha^2+4\alpha+1)\ln\alpha$	$\frac{1}{7}$	$\dfrac{4}{\alpha}$

It is interesting to note that for $n > 2, C_n(\alpha)$ is not very far from its value at $\alpha = 1$, even when α is as large as 20. Thus for $3 \leq n \leq 6$ or even for larger n the approximation $C_n(\alpha) \sim 1/(n+1)$ may be satisfactory for most purposes.

e. *Bias of V^* for the case $n_1 = n_2 = 10$*

The calculation of Ave $\{V^*\}$ by the above method is considerably

TABLE V
FRACTIONAL CORRECTION TO $1/w$

α	$C_1(\alpha)$	$C_2(\alpha)$	$C_3(\alpha)$	$C_4(\alpha)$	$C_5(\alpha)$	$C_6(\alpha)$
1	.5000	.3333	.2500	.2000	.1667	.1429
2	.5754	.3645	.2641	.2061	.1686	.1424
3	.6906	.4083	.2820	.2126	.1694	.1402
4	.8056	.4480	.2963	.2163	.1683	.1367
5	.9146	.4823	.3070	.2179	.1659	.1326
6	1.0172	.5119	.3149	.2180	.1629	.1283
7	1.1137	.5376	.3207	.2171	.1595	.1240
8	1.2050	.5601	.3250	.2155	.1559	.1199
9	1.2917	.5801	.3281	.2135	.1523	.1159
10	1.3742	.5979	.3303	.2112	.1488	.1121
20	2.0492	.7095	.3297	.1858	.1193	.0841
50	3.3891	.8274	.2928	.1354	.0759	.0487
100	4.8847	.8899	.2486	.0970	.0487	.0292

more difficult than the calculation of $V(\bar{\mu})$. An alternative procedure which applies for any n is to expand $1/(1 + \alpha t)^2$ in a power series in $1/(1 + t)$. We thus obtain convergent series expansions for C and for Ave $\{V^*\}$. These series have the disadvantage of converging rather slowly when α is large. Calculations were made for the case $n_1 = n_2 = 10$ and are shown in Table VI. The quantity Ave $\{1/\hat{w}\}$ is also shown for comparison as $1/\hat{w}$ might be taken a priori as a reasonable estimate of $V(\bar{\mu})$.

It appears that for $n_1 = n_2 = 10$, the bias of the uncorrected variance estimate, $1/\hat{w}$, is in the neighborhood of 15% for small α and

TABLE VI
TRUE VARIANCE AND AVERAGE ESTIMATED VARIANCES OF $\bar{\mu}$; $1/w = 1$
($n_1 = n_2 = 10$)

α	True Variance $V(\bar{\mu})$	Average estimated variances		Biases percent error in	
		Ave $\{V^*\}$	Ave $\{1/\hat{w}\}$	Ave $\{V^*\}$	Ave $\{1/\hat{w}\}$
1	1.091	1.077	0.909	−1.3%	−16.7%
2	1.088	1.073	0.917	−1.3%	−15.7%
3	1.082	1.068	0.928	−1.3%	−14.3%
4	1.077	1.062	0.936	−1.4%	−13.1%
α large	$1 + 1/\alpha$	$1 + 7/10\alpha$	$1 - 1/2\alpha$	$-30/(1 + \alpha)\%$	$-150/(1 + \alpha)\%$

decreases rather slowly. The bias of V^* is under 2%. The amount by which the correction reduces the bias varies from about 90% for small α to 80% for large α.

6. *Summary*

In combining estimates of an unknown parameter it may be reasonable to assume that the individual estimates are unbiased, but are derived from populations with possibly different variances. This paper provides first order corrections to the estimated variance of a weighted mean when the weights are the reciprocals of the estimated variances of the individual estimates.

In addition to the general results, described in the introduction, a special investigation is made for the case in which only two estimates are to be combined. Upper and lower bounds for the variance of the weighted mean are determined. Exact formulas for the variance of the weighted mean are given for the case with degrees of freedom proportional to the weights and for the case of arbitrary weights having the same number of degrees of freedom when this is less than or equal to 6.

The biases of the corrected and uncorrected methods are compared for the case of two estimates with both weights based on 10 degrees of freedom. The correction reduces the maximum bias from 16% to under 2%.

ACKNOWLEDGEMENT

The writer would like to express his thanks to Professor John W. Tukey for suggesting this problem and for his advice and encouragement throughout the investigation.

BIBLIOGRAPHY

[1] Aspin, Alice A., "An examination and further development of a formula arising in the problem of comparing two mean values," *Biometrika 35*, 1948, pp. 88–96.

[2] Aspin, Alice A., "Tables for use in comparisons whose accuracy involves two variances separately estimated," with Appendix by B. L. Welch, *Biometrika 36*, 1949, pp. 290–296.

[3] Carroll, Sarah Porter, "Relative accuracy of weighting inversely as the estimated variance," Unpublished Master's thesis, University of North Carolina.

[4] Carroll, Sarah P. and Cochran, W. G., "Weighting inversely as the estimated variance," (To be published).

[5] Cochran, W. G., "Problems arising in the analysis of a series of similar experiments," *Jour. Roy. Stat. Soc., Supp. 4*, 1937, pp. 102–118.

[6] Cochran, W. G., "Testing a linear relation among variances," *Biometrics 7*, 1951, pp. 17–32.

[7] Meier, P., "Weighted means and lattice designs," Unpublished Doctoral thesis, Princeton University, 1951.
[8] Neyman, J. and Scott, Elizabeth L., "Consistent estimates based on partially consistent observations," *Econometrika 16*, 1948, pp. 1–32.
[9] Smith, H. Fairfield, "The problem of comparing the results of two experiments with unequal errors," *Council Sci. Ind. Res. Jour.* (Australia) *9*, 1936, pp. 211–212.
[10] Snedecor, George W., "The statistical part of the scientific method," *Annals of the New York Acad. of Sci., 52*, 1950, pp. 742–749.
[11] Welch, B. L., "The significance of the difference between two means when the population variances are unequal," *Biometrika 29*, 1938, pp. 330–362.
[12] Yates, F. and Cochran, W. G., "The analysis of groups of experiments," *Jour. Agri. Sci. 28*, 1938, pp. 556–580.

PROCESSING DATA FOR OUTLIERS[1]

W. J. Dixon

University of Oregon

1. *Introduction*

Every experimenter has at some time or other faced the problem of whether certain of his observations properly belong in his presentation of measurements obtained. He must decide whether these observations are valid. If they are not valid the experimenter will wish to discard them or at least treat his data in a manner which will minimize their effect on his conclusions. Frequently interest in this topic arises only in the final stages of data processing. It is the author's view that a consideration of this sort is more properly made at the recording stage or perhaps at the stage of preliminary processing.

This problem will be discussed in terms of the following general models. We assume that observations are independently drawn from a particular distribution or alternatively, we assume that an observation is occasionally obtained from some other population and that there is nothing in the experimental situation to indicate that this has happened except what may be inferred from the observational reading itself.[2]

We assume that if no extraneous observations occur, the observations (or some transformation of them, such as logs) follow a normal distribution. We shall also assume that the occasional extraneous observations are either from a population with a shifted mean or from a population with the same mean and a larger variance. These assumptions may not be completely realistic but procedures developed for these alternatives should be helpful.

If one is taking observations where either of these models apply there remain two distinct problems.

First, one may attempt to pick out the particular observation or observations which are from the different populations. One may be interested in this selection either to decide that something has gone wrong with the experimental procedure resulting in this observation (in which case he will not wish to include the result) or that this observation gives an indication of some unusual occurrence which the investigator may wish to explore further.

[1]This research sponsored by the Office of Naval Research.

[2]There is no attempt here to discuss the problem of rejecting observations statistically when there are known experimental conditions which make the observation suspect. For example, the dirty test tube or the rat that died of the wrong disease.

The second problem is not concerned with tagging the particular observation which is from a different population, but to obtain a procedure of analysis not appreciably affected by the presence of such observations. This second problem is of importance whenever one wishes to estimate the mean or variance of the basic distribution in a situation where unavoidable contamination occasionally occurs.

The first problem—tagging the particular observation—is of importance in looking for "gross errors" or mavericks, or the best or largest of several different products. Frequently the analysis of variance test for difference in means is used in the latter case. This is not really a very good procedure since many types of inequality of means have the same chance of being discovered. It should be noted that the power of the analysis of variance test decreases as more products are considered when testing in a situation of one product different from others which are all alike.

The problem of testing *particular* observations as outliers was discussed in reference (1). The power of numerous criteria was investigated and recommendations were made there for various circumstances.

This paper will concern itself primarily with the problem of contamination occurring according to the following model:

Outliers occur with a certain probability each time an observation is made. Let $N(\mu, \sigma^2)$ represent a normal population with mean, μ, and variance, σ^2. An observation from $N(\mu + \lambda\sigma, \sigma^2)$ introduced into a sample from $N(\mu, \sigma^2)$ is termed a location error. An observation from $N(\mu, \lambda^2\sigma^2)$ introduced into a sample from $N(\mu, \sigma^2)$ is a scalar error. It will be convenient to use the notation $C_+(N, \gamma, \lambda)$ or $C_\times(N, \gamma, \lambda)$ to represent samples of size N drawn from a population $N(\mu, \sigma^2)$ contaminated γ proportion from $N(\mu + \lambda\sigma, \sigma^2)$ or from $N(\mu, \lambda^2\sigma^2)$, respectively.

Section 2 will discuss the estimation of μ by use of the mean and median. Section 3 discusses the estimation of σ and σ^2 by the sample variance and the range. Section 4 gives recommended rules for processing data under various conditions of contamination.

2. *Effects of Contamination on the Mean and Median.*

The median has often been proposed as an estimator for μ under certain conditions of contamination. The ability of the mean and median to estimate μ can be compared by computing the mean square error (MSE) of the estimates for various types of contamination. The biases will be listed in several cases. The bias of the arithmetic mean is defined as $E(\bar{x} - \mu)/\sigma$ and the MSE is defined as $E(\bar{x} - \mu)^2/\sigma^2$. The criteria of *better* estimate of mean to be used here is *smaller* MSE.

TABLE I. SAMPLES NOT TREATED FOR CONTAMINATION

	$C_+(5, .10, \lambda)$				$C_+(5, .01, \lambda)$				
	Mean		Median			Mean		Median	
λ	Bias	MSE	Bias	MSE	λ	Bias	MSE	Bias	MSE
0	0	.200	0	.287	0	0	.200	0	.287
2	.2	.313	.15	.41	2	.02	.208	.02	.30
3	.3	.455	.18	.48	3	.03	.219	.02	.30
5	.5	.908	.20	.61	5	.05	.252	.02	.30
7	.7	1.588	.22	.80	7	.07	.302	.02	.30

From Table I, it can be concluded that the median is superior to the mean of untreated data for 10% contamination in samples of size 5, only

Contours indicating equality of MSE of median of untreated data and MSE of $\bar{x}$ of treated data.

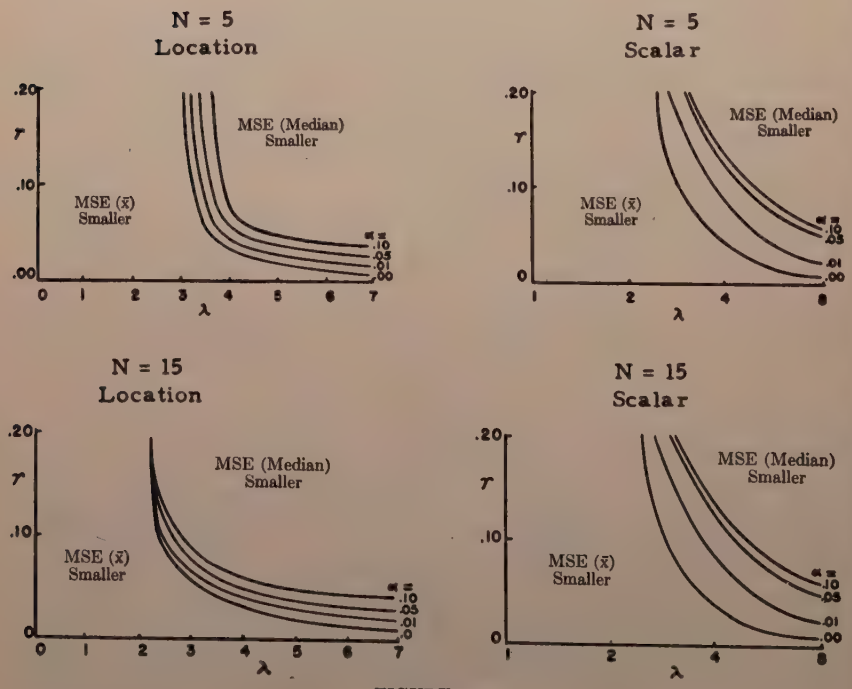


FIGURE 1.

if contamination is centered about 3.3σ or further from the mean. For 1% contamination the untreated mean is superior to the median for λ as large as 7. The reader is referred to Section 5 for the accuracy and method of obtaining the values in the above and succeeding tables.

The curves labelled $\alpha = 0$ in Figure 1 show the frequency and extent of contamination which can be tolerated before the MSE of the mean exceeds the MSE of the median. Curves are given for samples of size 5 and 15 and for location and scalar contamination. For example, for samples of size 5 which are 5% contaminated, the MSE, for $\bar{x}$ is smaller when the contaminating distribution is shifted 3σ but not when it is shifted 4σ.

Let us now consider changes in the above results when some of the contamination has been removed by the use of one of the r criteria of reference 2. A selection of critical values for these criteria is given in the Appendix.

Investigation was made using the 1, 5, and 10% levels of significance. The sample was tested until no further observations could be removed i.e. if a rejection was obtained at a certain level of significance, the reduced sample was again tested for outlier using the same level for α. This means, of course, that α should no longer be called a level of significance.

The additional curves in Figure 1 indicate the larger regions in which the MSE of $\bar{x}$ for treated samples is smaller than the MSE of the median for untreated samples when larger values of α are used. For extreme contamination an $\alpha = .20$ or .30 would further reduce the MSE for $\bar{x}$. This was not investigated in detail but it is known that this would not materially increase the size of λ and α which can be tolerated before the median should be used in preference to the mean.

In samples of size 5 use of the mean for treated samples results in most cases in a MSE *considerably smaller* than the MSE for the median. In cases of extreme contamination where the MSE for the median is smaller it is only slightly smaller. However, the MSE for the mean or the median is very large for heavy contamination. The use of the best treatment procedure still does not give us all that might be hoped for since the MSE is still large. The ratio of MSE for mean of treated data to MSE of mean for data with no contamination is an index of the extent of the contamination. We can also see from this index how much better the better estimate is. These ratios are given in Table II for samples of size 5 and 15. For samples of size 15 the picture is changed since the MSE for the mean becomes very large in the region where the MSE for the median is less than the MSE for the mean. Treatment is at the level $\alpha = .00, .01, .05$, or .10 which gives minimum MSE of $\bar{x}$.

TABLE II

$$\frac{\text{MSE}[\bar{x}, \text{ treated data}, C_+(5, \gamma, \lambda)]}{\text{MSE}[\bar{x}, N(\mu, \sigma^2)]}$$

γ \ λ	0	2	3	5	7
.01	1.00	1.04	1.08	1.2	1.3
.05	1.00	1.3	1.5	1.8	2.2
.10	1.00	1.6	2.2	3.4	5.1
.20	1.00	2.4	4.0	8.5	15.

$$\frac{\text{MSE [median, } C_+(5, \gamma, \lambda)]}{\text{MSE}[\bar{x}, N(\mu, \sigma^2)]}$$

γ \ λ	0	2	3	5	7
.01	1.43	1.5	1.5	1.5	1.5
.05	1.43	1.8	1.7	1.9	2.0
.10	1.43	2.1	2.4	3.1	4.0
.20	1.43	3.0	4.4	8.2	14.1

$$\frac{\text{MSE } [\bar{x}, \text{ treated data}, C_+(15, \gamma, \lambda)]}{\text{MSE}[\bar{x}, N(\mu, \sigma^2)]}$$

γ \ λ	0	2	3	5	7
.01	1.0	1.1	1.1	1.1	1.1
.05	1.0	1.3	1.7	1.9	2.3
.10	1.0	1.9	2.9	5.1	8.2
.20	1.0	4.0	7.6	20.	35.

$$\frac{\text{MSE [median, } C_+(15, \gamma, \lambda)]}{\text{MSE}[\bar{x}, N(\mu, \sigma^2)]}$$

γ \ λ	0	2	3	5	7
.01	1.6	1.6	1.6	1.6	1.6
.05	1.6	1.8	1.8	1.8	1.8
.10	1.6	2.3	2.3	2.3	2.3
.20	1.6	4.3	5.0	6.	7.

From these ratios we can see that in samples of size 5 even with extreme contamination the median is not a satisfactory substitute for $\bar{x}$ if the samples are treated for contamination. However, in samples of size 15 one should use the median if contamination beyond the $\alpha = .10$ curve (Figure 1) is expected.

The increase in MSE for $\bar{x}$ caused by removing occasional values which are not contaminators is very small in comparison to the reduction of MSE of $\bar{x}$ obtained by the removal of extreme contaminators.

Use of a large value of α will discover more contaminators but, of course, will increase the MSE of $\bar{x}$ if samples do not contain outliers. This effect is, however, small. Use of $\alpha = .10$ in samples of size 5 containing no contamination will increase MSE of $\bar{x}$ from .200 to approximately .216. Therefore, unless the contamination is believed to be slight a fairly large α should be used. In order to obtain minimum MSE for the estimate of μ, we can consider using one of the following procedures:

a) use of $\bar{x}$ after treating for rejection with $\alpha = .01$.
b) use of $\bar{x}$ after treating for rejection with $\alpha = .05$.
c) use of $\bar{x}$ after treating for rejection with $\alpha = .10$.
d) use of the median.

Table III gives the procedure resulting in the least MSE among the four procedures considered for the various types of contamination and sample sizes. The numbers in parentheses are the MSE resulting.

The MSE figures in Table III would not be increased by more than 5% by the use of $\alpha = .10$ in place of $\alpha = .01$ or .05 (or by 10% over no treatment). This is a small effect compared to the 50% to 200% increase in MSE resulting if certain extreme contamination is not removed. This fact will be taken into account in laying down general rules.

3. *Bias of s^2 and the range.*

The effect of contamination on the estimate of variance can be assessed by computing the amount of bias resulting. Let us define $B = E(s^2)/\sigma^2$. Since removal of outliers will reduce the variance, it is possible to make $B = 1$ for any γ and λ by choosing α sufficiently large. Since investigation was carried out only for $\alpha = .01, .05, .10$ (and a few results for $\alpha = .20$) it will not be possible to state the appropriate α for heavy or extreme contamination. Table IV lists the values obtained. It is believed that the consideration of bias alone is sufficient for the estimate of variance since in general the mean and MSE of the variance are closely related.

TABLE III. MINIMUM MSE FOR FOUR TREATMENTS

$N = 5$, location contamination

γ \ λ	0	2	3	5	7
.01	use $\bar{x}$ (.200)	a(.21)	a(.22)	c(.23)	c(.23)
.02		a(.22)	b(.24)	c(.26)	c(.27)
.05		b(.25)	b(.30)	c(.37)	c(.44)
.10		b(.32)	c(.43)	d(.62)	d(.80)
.20		b(.49)	c(.81)	d(1.63)	d(2.82)

$N = 15$, location contamination

γ \ λ	0	2	3	5	7
.01	use $\bar{x}$ (.067)	a(.07)	a(.07)	b(.07)	a(.07)
.02		a(.07)	a(.07)	c(.08)	b(.07)
.05		b(.08)	b(.11)	c(.13)	d(.12)
.10		b(.13)	d(.16)	d(.16)	d(.16)
.20		b(.27)	d(.33)	d(.4)	d(.4)

$N = 5$, scalar contamination

γ \ λ	1	2	4	8
.01	use $\bar{x}$ (.200)	no treat. (.21)	a(.22)	b(.24)
.02		a(.21)	b(.23)	b(.26)
.05		b(.23)	c(.26)	c(.36)
.10		b(.26)	c(.34)	d(.50)
.20		b(.31)	d(.49)	d(1.0)

$N = 15$, scalar contamination

γ \ λ	1	2	4	8
.01	use $\bar{x}$ (.067)	a(.07)	a(.07)	a(.07)
.02		a(.07)	a(.07)	a(.07)
.05		b(.07)	b(.08)	b(.10)
.10		b(.08)	c(.11)	c(.24)
.20		c(.08)	c(.23)	d(.70)

TABLE IV. APPROPRIATE α TO REMOVE BIAS IN s^2

$N = 5$, Location errors

γ \ λ	2	3	5	7
.01	.03	.05	.07	.05
.02	.06	.08	.10	.12
.05	.12			
.10				
.20				

$N = 15$, Location errors

γ \ λ	2	3	5	7
.01	.12	.10	.08	.01
.02	.14	.12	.10	.02
.05	.20	.25		
.10				
.20				

$N = 5$, Scalar errors

γ \ λ	2	4	8
.01	.02	.04	.05
.02	.03	.07	.10
.05	.07	.12	
.10	.12		
.20			

$N = 15$, Scalar errors

γ \ λ	2	4	8
.01	.10	.10	.05
.02	.20	.20	.10
.05			
.10			
.20			

TABLE V

Bias in s^2 for $C_+(5, \gamma, 5)$.

α \ γ	0	.10	.20
.00	1.00	3.52	6.0
.01	.99	1.90	5.3
.05	.95	1.52	4.4
.10	.91	1.26	4.0

The notation $\alpha = .00$ indicates results for untreated data.

Here again it is much more serious to allow contamination to remain than to remove non-contaminators incorrectly so that in general one should lean toward a large α. Table V illustrates this effect. For example, if contaminatoin is not present and we use $\alpha = .10$, we underestimate

TABLE VI

Appropriate α to remove bias of range estimate of σ.

Location errors

γ \ λ	2	3	5	7
.01	.01	.03	.04	.04
.02	.02	.05	.06	.05
.05	.06	.09	.10	.12
.10	.10			
.20				

Scalar errors

γ \ λ	2	4	8
.01	$<.01$	.01	.05
.02	.01	.04	.09
.05	.03	.05	
.10	.09		
.20			

σ^2 by 9%; but if 10% contamination at 5σ is present the use of the same rejection criterion will give us an overestimate of only 26% in place of 250% in samples of size 5.

In very small samples the *range* is often used to estimate the population standard deviation. Contamination will, of course, seriously affect the sample range, but the rejection criteria can effectively remove the bias in the range estimate of the population standard deviation for samples of size 5. Table VI shows the α which will result in an unbiased range estimate of σ in samples of size 5.

Table VII shows the bias of the range estimate of σ for one type of contamination. As before, it is much more serious to leave contamination in than to remove a few observations from samples which contain no contamination.

TABLE VII

Bias of the range estimate of σ for $C_+(5, \gamma, 5)$.

α \ γ	0	.10	.20
.00	1.00	1.66	2.11
.01	.99	1.50	1.91
.05	.97	1.27	1.63
.10	.93	1.13	1.48

The above results indicate that the range estimate of σ is less affected by contamination than s^2 even if no treatment is applied. Table VIII has been constructed to compare:

(1) s^2, the estimate of σ^2.
(2) ks, the estimate of σ where $E(ks) = \sigma$.
(3) the range estimate of σ.

TABLE VIII

Appropriate α to remove bias in $C_+(5, \gamma, 2)$.

γ	s^2	range estimate	ks
.01	.02	.01	.02
.02	.03	.02	.04
.05	.07	.06	.10
.10	.12	.10	$>.10$

The comparison is given in terms of the level α necessary to remove the bias in each of the estimates. The bias may be removed from the range estimate with a smaller value of α.

4. *Recommended Rules for Processing Data for Outliers.*

The problem of test of significance for tagging an individual as extraneous, extreme or as a "gross error" is pretty straightforward. We choose a level of significance, using the standard considerations and make a test on the set of observations we are processing. If a significant ratio is obtained we declare the extreme value to be from a population differing from that of the remaining observations. Depending on the practical situation we then declare the extreme individual a "gross erorr" or an exceptional individual. The best* statistic for this test if σ is known is the range over σ for outliers in either direction or the ratio $(x_n - \bar{x})/\sigma$ for a one-sided test. x_n represents the largest observation. For a one-sided test in the other direction we substitute $\bar{x} - x_1$ for $x_n - \bar{x}$. Here x_1 represents the smallest observation. The power of these tests is discussed in reference [1]. Critical values for range over σ are given in reference [4] and for $(x_n - \bar{x})/\sigma$ in reference [3].

If an independent estimate of σ is available, the best tests for outliers are the same as above with s replacing σ. Critical values for these tests are in references [3] and [4]. If no external estimate of σ is available the best statistics are the r-ratios of reference [2]. Critical values for these ratios are given in the Appendix.

Now, suppose that in place of tagging an individual observation from some different distribution, we wish to estimate the parameters of the basic distribution free from these contaminating effects. How might we process the data to come closer to the mean and variance of this basic distribution?

If very little is known about the contamination to be expected, about the best one can do is to "tag" observations as above and remove them from estimates of μ and σ.

If even a moderate amount of information about the type of contamination to be expected is available, a process can be prescribed which will minimize the effects of contamination on the estimates of mean and dispersion in small samples. The following rules result from the investigation of sections 2 and 3 for samples of size 5 and 15. Rules will be presented for these two sample sizes with the expectation that rules for samples of approximately these sizes will be approximately the same.

An attempt has been made to present simple rules for the estimation

*Best is used here in the sense of power greater than or equal to all other tests investigated in reference [1].

of both μ and σ. As a consequence the minimum MSE will not always be obtained. In most cases, however, the suggested procedure will yield a MSE which is not more than 5% larger than the minimum MSE. The rules will yield bias B between .90 and 1.10 for the indicated estimate of dispersion except in cases noted specifically in the rules.

Rules:

Process data using rejection criteria from Appendix I unless use of median is indicated. The appropriate α will be indicated in the rules below. Repeat application of criteria until no further observations are rejected. Use level of α as indicated in following statements.

$N = 5$, *Location contamination*

1. if $\gamma\lambda < .10$, use $\alpha = \gamma\lambda$, $\bar{x}$ for average, either s^2 or range for dispersion.
2. if $.10 < \gamma\lambda < .45$, use $\alpha = .10$, $\bar{x}$ for average, range for dispersion.
3. if $\gamma\lambda > .45$ use median for average, use $\alpha = .10$ and range to estimate dispersion. The estimate of dispersion will be biased, giving an overestimation of σ of more than 10%. ($B \simeq 1.1$ for $\gamma\lambda = .45$, $B \simeq 1.5$ for $\gamma\lambda = 1.00$).

$N = 5$, *Scalar contamination*

4. if $\gamma\lambda < .45$, use $\alpha = \frac{1}{2}\gamma\lambda$, $\bar{x}$ for average, s^2 for dispersion (for $\gamma\lambda > .30$ use range for dispersion. Bias for both range and s^2 over 10%).
5. if $\gamma\lambda > .45$, use median for average, range for dispersion. The estimate of dispersion will be biased, giving an overestimate of σ of more than 40%.

$N = 15$, *Location contamination*

6. if $\gamma\lambda < .30$, use $\alpha = .10$, $\bar{x}$ for average, s^2 for dispersion. For $\gamma > .02$ the estimate of dispersion will have considerable bias. ($B \simeq 1.2$ for $\gamma\lambda = .2$; $B \simeq 1.4$ for $\gamma\lambda = .3$).
7. if $\gamma\lambda > .30$ use the median for average, use $\alpha = .10$ and s^2 for dispersion. s^2 is considerably biased ($B \simeq 2$ for $\gamma\lambda = .50$).

$N = 15$, *Scalar contamination*

8. if $\gamma\lambda < 1.00$, use $\alpha = .10$, $\bar{x}$ for average, s^2 for dispersion. For $\gamma > .02$ (unless $\gamma\lambda < .15$) the estimate of dispersion will have considerable bias. ($B \simeq 1.1$ for $\gamma\lambda = .2$; $B \simeq 1.5$ for $\gamma\lambda = .4$).

9. if $\gamma\lambda > 1.00$ use median for average and use $\alpha = .10$ and s^2 for dispersion. s^2 is considerably biased ($B \simeq 10$ for $\gamma\lambda = 1.6$).

An application of the above rules is given as Example 1.

Example 1. Suppose samples of size 5 are taken from each lot. It is expected that about 10% of the observations will be location errors of 3 to 4 standard deviations. Here $\gamma = .10$ and $\lambda = 3$ to 4. Then $\gamma\lambda = .30$ to .40 and we use rule 2. Observations are recorded in order of size of measurement for a sample of five and treatment process indicated.

$x_1 = 23.2$ For $N = 5$ and $\alpha = .10$, the critical value of $r_{10} = .557$

$x_2 = 23.4$ By inspection $x_1 = 23.2$ is acceptable

$x_3 = 23.5$ The test for $x_5 = 25.5$ is

$x_4 = 24.1$ $$r_{10} = \frac{25.5 - 24.1}{25.5 - 23.2} = \frac{1.4}{2.3} = .609$$

$x_5 = 25.5$ x_5 is rejected

For $N = 4$ and $\alpha = .10$, the critical value of $r_{10} = .679$

The test for x_4 is

$$r_{10} = \frac{24.1 - 23.5}{24.1 - 23.2} = \frac{.6}{.9} = .667$$

x_4 is accepted.

The average is $(23.2 + 23.4 + 23.5 + 24.1)/4 = 23.55$.

The range is $24.4 - 23.2 = 0.9$.

The estimate of standard deviation is $(.486)(0.9) = .44$.

It may not be possible to decide which type of contamination might be expected in a particular sampling situation. Also the type of contamination might not be of the comparatively simple sort discussed here. However, if a large number of observations (say 50 to 100) are collected we can estimate the amount and type of contamination present. If we are willing to assume that the observations can be considered to be drawn from a population composed of two normal populations in different proportions and with possibly different means and variances there is a method for estimating these components. The estimation may be done by trial and error or graphically as described in References [5] and [6] and then one of the simpler models discussed here may be selected as representing approximately the actual conditions of contamination. A

trial and error method will be preferable if the population is not *very closely* represented by two normal populations.

Example 2. A series of chemical determinations are made on known chemical solutions giving a distribution as follows:

Error	Frequency		Fitted curve is:
	Observed	Fitted	
>1.25	2	2	
.9	6	5	80 *percent* $\mu = -.175$
.3	59	60	$\sigma = .333$
−.3	142	142	
−.9	38	39	
−1.5	11	9	20 *percent* $\mu = -.55$
−2.1	0	4	$\sigma = 1.0$
−2.7	2	2	
<−2.95	3	0	
	263		

There is a shift in mean of slightly more than one standard deviation unit. The important factor of contamination here is the large standard deviation of the second distribution. Considering only the scalar contamination, we have $\gamma = .20$ and $\lambda = 3$ or $\gamma\lambda = .6$. In samples of size 5 from the above population, the rules suggest use of the median and range.

The Appendix gives critical values for criteria for processing contaminated data and a table of multipliers for estimating the standard deviation from the range.

5. *Accuracy of Tabular Values.*

Although some known results are included and some were determined analytically, most results were obtained by sampling methods. Most of the sampling results for $N = 5$ are based on 100 samples and those for $N = 15$ on 66 samples. However, since the results quoted are weighted sums of several determinations each based on 100 (for $N = 5$) the effective sample size is greater than 100. Furthermore, sampling results were obtained for several values of a parameter (e.g. λ) so that unfortunately large sampling deviations could in some cases be discovered and rectified by an increased amount of sampling. It is difficult to state the accuracy

to be associated with each figure, but the accuracy should be adequate for determining the comparatively large differences on which the recommended analysis is based. After the tables had been assembled several of the reported results were checked by additional sampling. No MSE or bias differed from the tabulated results by more than 15%. Errors of 15 or 20% would not change the recommended procedures appreciably.

The values known to be correct are all results reported for $\lambda = 0$, the quantities for the mean in Table I, the results for no contamination in Table II and III and the first line of Table V.

REFERENCES

1. W. J. Dixon, "Analysis of Extreme Values," *Annals of Math. Stat.*, Vol. 21 (1950) pp. 488–506.
2. W. J. Dixon, "Ratios Involving Extreme Values," *Annals of Math. Stat.*, Vol. 22 (1951) pp. 68–78.
3. K. R. Nair, "Tables of Percentage Points of the 'Studentized' Extreme Deviate From the Sample Mean," *Biometrika*, Vol. 39 (1952) pp. 189–191.
4. Joyce M. May, "Extended and Corrected Tables of the Upper Percentage Points of the 'Studentized' Range," *Biometrika*, Vol. 39 (1952) pp. 192–193.
5. Carl Burrau, "The Half-Invariants of the Sum of Two Typical Laws of Errors, with an Application to the Problem of Dissecting a Frequency Curve into Components," *Skandinavisk Aktuarietidskrift*, Vol. 17 (1934), pp. 1–6.
6. Bengt Stromgren, "Tables and Diagrams for Dissecting a Frequency Curve into Components by the Half-invariant Method," *Skandinavisk Aktuarietidskrift*, Vol. 17 (1934), pp. 7–54.

APPENDIX

CRITICAL VALUES AND CRITERIA FOR TESTING FOR EXTREME VALUES

N \ α	.30	.20	.10	.05	.02	.01	.005	Criterion
3	.684	.781	.886	.941	.976	.988	.994	
4	.471	.560	.679	.765	.846	.889	.926	
5	.373	.451	.557	.642	.729	.780	.821	$r_{10} = \frac{x_N - x_{N-1}}{x_N - x_1}$
6	.318	.386	.482	.560	.644	.698	.740	
7	.281	.344	.434	.507	.586	.637	.680	
8	.318	.385	.479	.554	.631	.683	.725	
9	.288	.352	.441	.512	.587	.635	.677	$r_{11} = \frac{x_N - x_{N-1}}{x_N - x_2}$
10	.265	.325	.409	.477	.551	.597	.639	
11	.391	.442	.517	.576	.638	.679	.713	
12	.370	.419	.490	.546	.605	.642	.675	$r_{21} = \frac{x_N - x_{N-2}}{x_N - x_2}$
13	.351	.399	.467	.521	.578	.615	.649	
14	.370	.421	.492	.546	.602	.641	.674	
15	.353	.402	.472	.525	.579	.616	.647	
16	.338	.386	.454	.507	.559	.595	.624	
17	.325	.373	.438	.490	.542	.577	.605	
18	.314	.361	.424	.475	.527	.561	.589	
19	.304	.350	.412	.462	.514	.547	.575	
20	.295	.340	.401	.450	.502	.535	.562	$r_{22} = \frac{x_N - x_{N-2}}{x_N - x_3}$
21	.287	.331	.391	.440	.491	.524	.551	
22	.280	.323	.382	.430	.481	.514	.541	
23	.274	.316	.374	.421	.472	.505	.532	
24	.268	.310	.367	.413	.464	.497	.524	
25	.262	.304	.360	.406	.457	.489	.516	

RANGE ESTIMATE OF STANDARD DEVIATION WHERE SAMPLE RANGE = w.

N	Estimate
2	.886w
3	.591w
4	.486w
5	.430w
6	.395w
7	.370w
8	.351w
9	.337w
10	.325w

THE ESTIMATION OF HERITABILITY BY REGRESSION OF OFFSPRING ON PARENT[1]

O. Kempthorne and O. B. Tandon

Iowa State College

I. *Introduction*

The concept of heritability is associated with the relative importance of heredity and of environment as they influence the variations in a character. Knowing the degree of heritability of a characteristic is very helpful in choosing an efficient breeding system, in estimating the gain to be expected under mass selection and in constructing a selection index [see for example Hazel (1943)].

Methods of estimating heritability all depend upon the degree to which related animals resemble each other more than less closely related animals do. One of the most useful methods is based on the resemblance between parents and offspring. In general this method is less likely to have been seriously affected by environmental contributions than are estimates based on the resemblance of two contemporary relatives or the resemblance of two maternal sibs who have had a common prenatal environment. Also the sampling errors of the estimates obtained by comparing parent and offspring are likely to be less serious than those of estimates based on individuals less closely related. Furthermore the estimates of degree of heritability based on parent and offspring do not include dominance deviations as do the ones based on full sibs.

The resemblance between parent and offspring could be measured in either of the two ways: by the regression of offspring on parent or by correlation between parent and offspring. Certain features which are common to most animal husbandry data, make the use of regression preferable to that of correlation for estimating heritability.

[1]Journal Paper No. J.2184 of the Iowa Agricultural Experiment Station, Project 890, Ames, Iowa.

Measuring the regression of offspring on parent is straight forward if the number of offspring from each parent is constant. However in most animal husbandry data the parents do not all have the same number of offspring. In computing the regression with variable number of offspring per parent the problem arises of how best to weight these numbers. Two practices have been widely used. One is to repeat the parent's record with each offspring's record. The other is to average all the offspring of a parent and regress each such average on the appropriate parent's record. The former practice would be valid if the correlation among the offspring of a parent were zero while the latter would be valid if the correlation among members of each progeny group were one. Obviously the real situation in most animal husbandry material is intermediate to these two extreme conditions, although usually nearer to the former.

The present analysis was undertaken to find the "best" procedure in the sense of finding what intermediate weighting system will give unbiased estimates of the regression, with minimum sampling variance. First to be considered will be the general case of estimating the regression coefficient when a variable number of observations have been taken on the dependent variable and only one observation has been made on the independent variable. Later some special cases will be considered.

Derivation of an estimator for the general case:

If two variables, x and y, have the following properties:

$$
\begin{aligned}
&E(x_i) = \mu \qquad E(y_{ij}) = \mu_y \\
&E(x_i - \mu)^2 = E(y_{ij} - \mu_y)^2 = \sigma_y^2 \\
&E(y_{ij} - \mu_y)(x_i - \mu) = \beta\sigma_y^2 \\
&E(y_{ij} - \mu_y)(y_{ij'} - \mu_y) = \rho_1\sigma_y^2 \\
&E(y_{im} - \mu_y)(y_{jk} - \mu_y) = E(x_i - \mu)(y_{jk} - \mu_y) \\
&\qquad\qquad\qquad\qquad = E(x_i - \mu)(x_j - \mu) \\
&\qquad\qquad\qquad\qquad = 0 \text{ when } i \neq j
\end{aligned}
$$

then we may obtain the regression of y on x as follows:

$$y_{ij} = \mu_y + \beta(x_i - \mu) + e_{ij}$$

where y_{ij} is the jth observation on the character y in a group for which the observed value of the character x is x_i. The x's are assumed to be measured without error. μ is an effect common to all the x's and μ_y is

an effect common to all the y's; β is the regression of y's on the x's; e_{ij} is the residual peculiar to the jth y in the ith group. Let us examine the distributional properties of the e_{ij}'s:

$$E(e_{ij}) = 0$$

$$E(e_{ij})^2 = E[y_{ij} - \mu_y - \beta(x_i - \mu)]^2$$

$$= \sigma_y^2[1 - \beta^2], \text{ which will be denoted by } \sigma^2.$$

Also,

$$E(e_{ij}e_{ij'}) = E[\{y_{ij} - \mu_y - \beta(x_i - \mu)\}\{y_{ij'} - \mu_y - \beta(x_i - \mu)\}]$$

$$= \rho_1\sigma_y^2 - \beta^2\sigma_y^2$$

$$= \sigma_y^2[\rho_1 - \beta^2], \text{ which will be denoted by } \rho\sigma^2.$$

ρ then is the correlation coefficient between e_{ij} and $e_{ij'}$ and it equals

$$\frac{E(e_{ij}e_{ij'})}{E(e_{ij})^2} = \frac{\rho_1 - \beta^2}{1 - \beta^2}$$

Using the model given earlier now consider $L = \Sigma_{ij}\, \lambda_{ij}y_{ij}$ as an estimator of β. L will then be a linear function of the y_{ij}'s with λ_{ij}'s being used as the weights. We will now proceed to obtain an expression for these λ_{ij}'s. For L to be unbiased we need

$$\sum_{ij} \lambda_{ij} = 0 \qquad \text{and}$$

$$\sum_{ij} \lambda_{ij}x_i = 1$$

From all such unbiased estimators we wish to choose the one that has the minimum variance.

Now the variance of L is

$$\sigma^2(1 - \rho) \sum_{ij} \lambda_{ij}^2 + \sigma^2\rho \sum_i \left(\sum_j \lambda_{ij}\right)^2$$

$$= \sigma^2(1 - \rho)\left[\sum_{ij} \lambda_{ij}^2 + T \sum_i \left(\sum_j \lambda_{ij}\right)^2\right],$$

where

$$T = \frac{\rho}{1 - \rho} = \frac{\rho_1 - \beta^2}{1 - \rho_1} \tag{1}$$

It is clear that the variance of the defined linear estimator of β depends on the true value of β and ρ_1 since this variance depends on T which is the function given above (1) of β and ρ_1.

Let us suppose that we can guess the value of T in the population we are investigating and denote the guessed value by τ. It is reasonable to

consider the linear unbiased estimator of β which has minimum variance if in fact the true value of T is τ. The estimator we shall obtain is not the best linear unbiased estimator but will be unbiased and will be close to the best linear unbiased estimator to the extent that we guess a value τ which is close to the true value T.

Following this approach then, and since we may ignore $\sigma^2(1 - \rho)$ which will be the same for all linear estimators of the type considered, we shall minimize,

$$\sum_{ij} \lambda_{ij}^2 + \tau \sum_i \left(\sum_j \lambda_{ij}\right)^2 \qquad \text{subject to the conditions}$$

$$\sum_{ij} \lambda_{ij} = 0 \qquad \text{and} \qquad \sum_{ij} \lambda_{ij} x_i = 1.$$

Using 2π and 2ϕ as Lagrange multipliers, we therefore take an unconditional minimum of

$$Q = \sum_{ij} \lambda_{ij}^2 + \tau \sum_i \left(\sum_j \lambda_{ij}\right)^2 - 2\pi \sum_{ij} \lambda_{ij} - 2\phi \sum_{ij} \lambda_{ij} x_i .$$

and utilizing the conditions we get

$$\lambda_{ij} = \lambda_i = \frac{1}{1 + n_i\tau} \left(\frac{x_i - \bar{x}}{\sum_i \frac{n_i}{1 + n_i\tau} (x_i - \bar{x})^2} \right),$$

where

$$\bar{x} = \frac{\sum_i \frac{n_i}{1 + n_i\tau} x_i}{\sum_i \frac{n_i}{1 + n_i\tau}}$$

As a result of this process, we have obtained a linear unbiased estimator of β, namely

$$\sum_{ij} \lambda_{ij} y_{ij} = \sum_i n_i \lambda_i y_{i.} = \sum_i \frac{n_i}{1 + n_i\tau} \left(\frac{x_i - \bar{x}}{\sum_i \frac{n_i}{1 + n_i\tau} (x_i - \bar{x})^2} \right) y_{i.} ,$$

where

$$y_{i.} = \frac{1}{n_i} \sum_j y_{ij}$$

and this we denote by $\hat{\beta}$.
If we let $w_i = n_i/(1 + n_i\tau)$ then the estimator is

$$\frac{\sum_i w_i (x_i - \bar{x}) y_{i.}}{\sum_i w_i (x_i - \bar{x})^2} ,$$

and the variance of the estimator is therefore

$$\sigma^2(1-\rho)\sum_i n_i \frac{(1+n_iT)}{(1+n_i\tau)^2}\frac{(x_i-\bar{x})^2}{\{\sum_i w_i(x_i-\bar{x})^2\}^2} \tag{2}$$

We may note that if τ in fact equals T this reduces to

$$\sigma^2(1-\rho)\frac{\sum_i w_i(x_i-\bar{x})^2}{[\sum_i w_i(x_i-\bar{x})^2]^2} = \frac{\sigma^2(1-\rho)}{\sum_i w_i(x_i-\bar{x})^2}$$

The estimated variance of the estimate:

The variance of the estimate involves T as is shown in (2). To estimate the variance of the estimate we have obtained, we may substitute in (2) an estimate $\hat{T}$, of T. We may obtain $\hat{T}$ from an estimate of ρ_1 and the estimate $\hat{\beta}$. It should be noted here that the use of $\hat{T}$, thus obtained, will not give an unbiased estimate of the multiplier of $\sigma^2(1-\rho)$ in the variance because unbiasedness in $\hat{\beta}$ will not result in unbiasedness in

$$\hat{T} = \frac{\hat{\rho}_1 - \hat{\beta}^2}{1-\hat{\rho}_1}$$

The amount of bias may however be expected to be small if the data are at all adequate to give an indication of β and ρ_1 .

There remains the question of the estimation of $\sigma^2(1-\rho)$. Consider the mean square between y's within x groups, i.e.

$$\frac{\sum_{ij}(y_{ij}-y_{i\cdot})^2}{\sum_i (n_i-1)} = \frac{\sum_i\sum_j y_{ij}^2 - \sum_i \frac{(\sum_j y_{ij})^2}{n_i}}{\sum (n_i-1)}$$

The expectation of this quantity is

$$\frac{1}{\sum_i (n_i-1)}\sum_i\left[\sum_j E(y_{ij})^2 - \frac{1}{n_i}\{\sum_j E(y_{ij})^2 + \sum_{j\neq j'} E(y_{ij}y_{ij'})\}\right]$$

Since

$$E(y_{ij})^2 = \mu_y^2 + \sigma_y^2$$

and

$$E(y_{ij}y_{ij'}) = \mu_y^2 + \rho_1\sigma_y^2$$

the expectation of the mean square between y's within x groups is

$$\frac{1}{\sum_i (n_i - 1)} \sum_i \Big[n_i(\mu_y^2 + \sigma_y^2) - \frac{1}{n_i} \{ n_i(\mu_y^2 + \sigma_y^2) + n_i(n_i - 1)(\mu_y^2 + \rho_1\sigma_y^2) \} \Big]$$

$$= \frac{1}{\sum_i (n_i - 1)} \sum_i [(n_i - 1)\sigma_y^2 - (n_i - 1)\rho_1\sigma_y^2]$$

$$= \sigma_y^2(1 - \rho_1)$$

which is equal to $\sigma^2(1 - \rho)$.

Summary of procedure:

I. Guess $T = \rho/(1 - \rho)$ and call the guessed value τ.

II. Estimate β by

$$\hat{\beta} = \frac{\sum_i w_i(x_i - \bar{x})y_{i\cdot}}{\sum_i w_i(x_i - \bar{x})^2} \qquad \text{where} \qquad w_i = \frac{n_i}{1 + n_i\tau} \qquad \text{and}$$

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \qquad \text{and } \tau \text{ is the guessed value of } T.$$

III. Estimate ρ_1 by using the mean squares within and between x groups. The expectation of the mean square within groups is $\sigma_y^2(1 - \rho_1)$, and of the mean square between groups is

$$\sigma_y^2(1 - \rho_1) + \frac{1}{k - 1}\Big(\sum n_i - \frac{\sum n_i^2}{\sum n_i}\Big)\sigma_y^2\rho_1 ,$$

where k is the number of x groups, so that one can estimate ρ_1 by equating observed to expected mean squares.

IV. Using the estimates $\hat{\rho}_1$ and $\hat{\beta}$, we obtain

$$\hat{T} = \frac{\hat{\rho}_1 - \hat{\beta}^2}{1 - \hat{\rho}_1}$$

V. The estimated variance of $\hat{\beta}$ is then the mean square among y's within x groups times

$$\sum_i n_i \frac{1 + n_i\hat{T}}{(1 + n_i\tau)^2} \frac{(x_i - \bar{x})^2}{\{\sum_i w_i(x_i - \bar{x})^2\}^2}$$

Extension of the method:

It is obvious that we may extend the method of estimation obtained above, so that the ultimate solution is obtained by iteration as is the case with the application of maximum likelihood to many situations. We could use the estimated value $\hat{T}$ as a value for τ and repeat the process until the estimated $\hat{T}$ does not change. This would be essentially the same as using the method of maximum likelihood, and would of course lead to a non-linear estimate, properties of which are very difficult to determine. By using the originally guessed value of T we obtain a linear unbiased estimate whereas we cannot claim that the solution obtained by iteration is unbiased.

The Regression of offspring on a parent: a special case

Before actually considering an application to this case of the results derived in the general case it seems appropriate that the following comments be made.

The Basic Genetic Model:

A complete historical review of the following concepts is not intended here. The idea of heritability, for example, dates back at least to the writings of some of the geneticists as early as the beginning of the twentieth century. The 1918 paper of R. A. Fisher appears to be the first paper in which general quantitative relationships between relatives were deduced. For this reason only here specific reference to that paper will be made.

In the 1918 paper Fisher showed that under certain assumptions the regression of offspring on a parent in a random mating population is given by the ratio $\sigma_g^2/2\sigma_p^2$ where σ_g^2 is the additively genetic variance defined in a least square sense and σ_p^2 is the total variance in the population. This is true if the effects of different loci in the population are combined in a strictly additive manner. Fisher further considered the effects of epistasis, homogamy, the presence of multiple alleles and of the presence of linkage on the relationships between relatives. In the absence of epistasis he showed that the correlation between half-sibs was $\sigma_g^2/4\sigma_p^2$ while the correlation between full-sibs was $\sigma_g^2/2\sigma_p^2$ plus a fraction of dominance variance. It is clear that the ratio σ_g^2/σ_p^2 is crucial in a description of correlation between relatives and also in prediction of breeding performance.

Let us denote by x's the phenotypic values of one of the parents and denote by y's the phenotypic values of offspring. We can, as indicated

earlier, obtain the regression of offspring on one of the parents. Now we will assume that the x's and the y's have the properties indicated earlier, then β, the regression of offspring's phenotype on the phenotype of the parent, in a random mating population without epistasis will equal

$$\frac{1}{2}\frac{\sigma_g^2}{\sigma_p^2}$$

It is not meant to imply here that the above relationships are exactly satisfied because epistasis may be of considerable importance. If epistasis is an important source of variation in the characteristic under study a fraction would have to be added to the regression above. This fraction would measure the extra relationship between the phenotypes of the parents and their offspring that is brought about because certain proportion of interallelic combinations are alike between a parent and its offspring. The regression coefficient in such a case will be larger than if epistasis were not involved.

Without any loss of generality we may consider that the parent whose phenotype is being used in the computation of the regression coefficient is the dam. If the regression were computed on an intra-sire basis it can be seen that it is not necessary to assume that the sires used constitute a random sample of the population of sires.

Denoting σ_g^2/σ_p^2 by h which is commonly called heritability in the narrow sense we note that the intra-sire regression of offspring's phenotype on the phenotype of the dam will equal half the heritability, if epistasis is negligible. In fact the parameter we would like to estimate is h. We shall do this however by estimating β and then doubling the estimate of β.

A very special case of the regression of offspring's phenotype on the phenotype of dam would be one where no two offspring from any one dam are full-sibs. This is not a very strong restriction as it may seem at first. In dairy cattle it is fairly common to find cases where full-sibs constitute five percent or less of all the offspring. An advantage of this special case is that one does not have to compute the regression within sire groups. In such a case the problem of knowing appropriate weights becomes of great importance because the proportion of dams with more than one offspring each becomes larger than the proportion of dams with more than one offspring each by a common sire. Now we will illustrate the method of estimating a regression coefficient for the case considered above. We will also compute the heritability from it.

An example:

The data for this example came from the Holstein herd at Iowa State College, Ames, Iowa, U.S.A. and include the production of milk in the first eight months of the first lactation of 133 cows and their 185 daughters. The data extended over a period of eleven years beginning in 1940 and ending in 1950.

The production records were corrected to mature equivalent by multiplying each record by a correction factor corresponding to the age at which the record was made. The age correction factors used for this study were the same as those given by Kendrick (1935) for Holstein cows. The analyses which follow were all made on the adjusted yields obtained after this age correction.

TABLE 1

No. of daughters per dam	$\Sigma x_i y_{i.}^*$	Σx_i^2	Σx_i	$\Sigma y_{i.}$	No. of dams
1	11,281,508,844	11,233,446,762	1,018,238	1,048,996	95
2	3,115,132,934	3,231,213,647	274,855	271,607	24
3	1,754,765,230	1,794,941,038	157,458	154,841	14

Using $\tau = .04$ as a guessed value of T, it is found that $w_1 = .9615$, $w_2 = 1.8519$ and $w_3 = 2.6786$. Applying these weights to the sums, squares and products the following cross product and the sum of squares around the means were obtained noting that w_i is constant within no. of daughter classes:

$$\sum w_i x_i y_{i.} - \frac{(\sum w_i x_i)(\sum w_i y_{i.})}{\sum w_i} = 86{,}091{,}894$$

$$\sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i} = 544{,}850{,}992$$

Thus

$$\hat{\beta} = \frac{86{,}091{,}894}{544{,}850{,}992} = .1583.$$

The estimate of ρ_1 in this case was -0.018. These lead to an estimate of

$$\hat{T} = \frac{\hat{\rho}_1 - \hat{\beta}^2}{1 - \hat{\rho}_1} = -0.042.$$

*$y_{i.}$ is the arithmetic average of maternal half-sibs. No full-sibs are included in this case.

The estimate of heritability obtained by doubling the estimate of the coefficient of regression of daughter's milk production on dam's milk production is then .3166. For comparison with the present estimate, other estimates of heritability were computed by repeating the dam's record with each daughter's record and also by regressing the averages of all the daughters of a dam on the dam's record without using any weights. The regression coefficients obtained thus and the corresponding heritabilities are given in Table 2.

TABLE 2

	$\hat{\beta}$	Heritability
Repeating the dam's record with each daughter's record	.160 ± .078	.321 ± .156
Weighted regression	.158 ± .078	.317 ± .156
Unweighted regression of means of daughters on dams	.136 ± .090	.271 ± .180

The formulae for estimating the variances of the three types of regression coefficients can be easily derived. They are:

$$\hat{\sigma}^2_{\hat{\beta}}\,(\text{dams repeated}) = \sigma^2(1 - \hat{\rho}) \sum_i n_i(1 + n_i\hat{T}) \frac{(x_i - \bar{\bar{x}})^2}{[\sum n_i(x_i - \bar{\bar{x}})^2]^2}$$

where n_i = number of daughters from the ith dam;

$$\bar{\bar{x}} = \frac{\sum n_i x_i}{\sum n_i};$$

$$\hat{\sigma}^2_{\hat{\beta}}\,(\text{weighted}) = \sigma^2(1 - \hat{\rho}) \sum_i n_i \frac{(1 + n_i\hat{T})}{(1 + n_i\tau)^2} \frac{(x_i - \bar{x})^2}{[\sum w_i(x_i - \bar{x})^2]^2}$$

where τ is the guessed value of T which was used to compute the weights, .04 in the present case;

$$w_i = \frac{n_i}{1 + n_i\tau} \qquad \text{and} \qquad \bar{x} = \frac{\sum w_i x_i}{\sum w_i};$$

$$\hat{\sigma}^2_{\hat{\beta}}\,(\text{unweighted}) = \sigma^2(1 - \hat{\rho}) \sum_i \frac{(1 + n_i\hat{T})}{n_i} \frac{(x_i - \bar{x})^2}{[\sum (x_i - \bar{x})^2]^2}$$

where $\bar{x} = \sum x_i/k$, k being the number of dams.

In order to estimate the variances of the estimates we have computed, an estimate of $\sigma^2(1 - \rho)$ was obtained by using the mean squares

among maternal half-sibs. The mean squares among offspring from different dams and the mean squares among maternal half-sibs in the present case were 3,659,692 and 3,752,699 respectively. The negative value of $\hat{\rho}_1$, which stems from the smallness of the former mean squares as compared to the latter, indicates that there is a negative correlation among all causes other than the dam which affect the milk production of material half-sibs. Since the value of $\hat{\rho}_1$ is not statistically significant, one might be lead to think that the present value is attributable to sampling variations and that the true value of ρ_1 in such a population is probably very small.

It may be noted that in the present instance there is little to choose between the three estimators considered. The smallness of the difference between the estimates arises because a large proportion of the records are from dams with only one offspring and because the value of $\hat{T}$ is very small. It should also be noted that estimates of error obtained by either of the alternative methods considered and ignoring the correlation between half-sibs will be biased downwards.

REFERENCES

1. Fisher, R. A. 1918. On the correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edin.* 52:399–433.
2. Hazel, L. N. 1943. The genetic basis for constructing selection indexes. *Genetics* 28:476–490.
3. Kendrick, J. F. 1941. Standardizing dairy herd improvement associations records in proving sires. *U.S.D.A. Bureau of Dairy Industry.* (*Mimeo. Circular 925*).

A NOTE ON RECTANGULAR LATTICES[1]

K. R. NAIR[2]

Institute of Statistics
University of North Carolina

I. *The Simple, Triple and Near Balance Rectangular Lattices.*

1. In a recent paper the author (1951) showed that the simple rectangular lattice for $p(p-1)$ varieties or treatments in blocks of $(p-1)$ plots developed by Harshbarger (1947) is a partially balanced incomplete block (p.b.i.b.) design and that the triple rectangular lattice developed by him (1949) for the same number of varieties and size of block is not a p.b.i.b. design, except when $p = 3$ and 4.

2. Harshbarger (1951) has recently developed the Near Balance Rectangular Lattice for $p(p-1)$ varieties in blocks of $(p-1)$ plots and with every variety replicated p times. It is interesting to note that this lattice is a p.b.i.b. design having two associate classes. In fact, it follows as a special case of a general rectangular lattice design given on p. 370 (Section 7.5) of the paper by Bose and Nair (1939). The parameters of their design are:

$$v = pq, \quad k = q, \quad r = p, \quad b = p^2 \quad (p > q)$$

$$\lambda_1 = 1 \quad \lambda_2 = 0$$

$$n_1 = p(q-1) \quad n_2 = (p-1)$$

$$p_{ik}^1 = \begin{Bmatrix} p(q-2) & (p-1) \\ (p-1) & 0 \end{Bmatrix}; \qquad p_{ik}^2 = \begin{Bmatrix} p(q-1) & 0 \\ 0 & (p-2) \end{Bmatrix}$$

$$\text{Efficiency factor} = \frac{(q-1)(pq-1)}{(q-1)(pq-1) + q(p-1)}$$

3. When $q = (p-1)$, the above design becomes the Near Balance Rectangular Lattice.

II. *The n-ple Rectangular Lattice*

1. Consider the n-ple rectangular lattice for $p(p-1)$ treatments or varieties in blocks of $(p-1)$ plots and with every treatment repli-

[1]Presented before a joint meeting of the Institute of Mathematical Statistics and The Biometric Society (ENAR) at Blacksburg, Virginia on 20 March, 1952.

[2]Present address: Forest Research Institute, Dehra Dun, India.

cated n times ($2 \leq n \leq p$). In the usual notation used for incomplete block designs:

$$v = p(p-1), \quad k = (p-1), \quad r = n, \quad b = np \quad (2 \leq n \leq p)$$

When $n = 2$, 3 and p respectively we have the simple, triple and near balance rectangular lattices.

2. A convenient method of constructing the n-ple rectangular lattice is to take a set of $(n-2)$ orthogonal squares of size p in which the elements in the leading diagonal of each square will all be different. Out of the $(p-1)$ orthogonal squares that can be constructed when p is a prime integer or power of a prime integer, only $(p-2)$ squares can be written in that form. By suitable interchange of the elements 1, 2, 3, $\cdots$ p of each square we can make the elements of the leading diagonals of these $(p-2)$ squares appear in the natural order 1, 2, 3, $\cdots$ p. Such squares are known as squares with *ordered directrices*. The $(p-2)$ squares with ordered directrices can easily be constructed for $p = 3, 4, 5, 7, 8$, and 9, from the complete sets of orthogonal squares given by Fisher and Yates (1948).

3. We have already seen that the n-ple rectangular lattice is a p.b.i.b. design when $n = 2$ and p. It can be shown that, when $n = (p-1)$, the n-ple rectangular lattice becomes a p.b.i.b. design with the following parameters:

$$v = b = p(p-1) \qquad r = k = (p-1)$$

$$\lambda_1 = 1 \qquad \lambda_2 = 0 \qquad \lambda_3 = 0$$

$$n_1 = (p-1)(p-2) \qquad n_2 = (p-1) \qquad n_3 = (p-2)$$

$$p_{ik}^1 = \begin{Bmatrix} (p-2)(p-3) & (p-2) & (p-3) \\ (p-2) & 0 & 1 \\ (p-3) & 1 & 0 \end{Bmatrix}$$

$$p_{ik}^2 = \begin{Bmatrix} (p-2)^2 & 0 & (p-2) \\ 0 & (p-2) & 0 \\ (p-2) & 0 & 0 \end{Bmatrix}$$

$$p_{ik}^3 = \begin{Bmatrix} (p-1)(p-3) & (p-1) & 0 \\ (p-1) & 0 & 0 \\ 0 & 0 & (p-3) \end{Bmatrix}$$

Efficiency factor

$$= \frac{p(p-2)(p^2-p-1)(p^2-3p+1)}{p(p-2)(p^2-p-1)(p^2-3p+1) + p(p-1)^2(p^2-3p+1) + (p-1)^3}$$

The second and third associates are distinguished from each other with the help of the fact that two treatments which are third associates will have all the $n = (p - 1)$ suffixes common. Pairs of treatments which have $(p - 2)$ suffixes common would have either occurred together in a block or not occurred. They form the first and second associate classes respectively.

Let us consider the case $n = 4$, $p = 5$ as an example. The 4 replications of the 20 treatments can be formed from the composite table which follows:

*	$a_1b_2c_4d_3$	$a_1b_3c_2d_5$	$a_1b_4c_5d_2$	$a_1b_5c_3d_4$
$a_2b_1c_4d_5$	*	$a_2b_3c_5d_4$	$a_2b_4c_3d_1$	$a_2b_5c_1d_3$
$a_3b_1c_2d_4$	$a_3b_2c_5d_1$	*	$a_3b_4c_1d_5$	$a_3b_5c_4d_2$
$a_4b_1c_5d_3$	$a_4b_2c_3d_5$	$a_4b_3c_1d_2$	*	$a_4b_5c_2d_1$
$a_5b_1c_3d_2$	$a_5b_2c_1d_4$	$a_5b_3c_4d_1$	$a_5b_4c_2d_3$	*

The first, second and third associates of any treatment can easily be written down. Thus, taking treatment $a_1b_2c_4d_3$ as an example, its associates are:

First	Second	Third
abcd	*abcd*	*abcd*
1325	2354	2431
1452	3415	3124
1534	4521	4312
3251	5132	
4235		
5214		
2145		
3542		
5341		
2513		
4153		
5423		

4. When $p = 5$ and if four replications are available it is better to use the design with $n = 4$ rather than duplicate the design with $n = 2$.

The efficiency factor in the former case is

$$\frac{3135}{4079} = 0.769$$

The efficiency factor for the simple rectangular lattice is

$$\frac{p(p-2)(p^2-p-1)}{p(p-2)(p^2-p-1)+2(p-1)^3+2(p-1)}$$

In our example of $p = 5$, $n = 2$, the efficiency factor is, therefore,

$$\frac{285}{421} = 0.677$$

Hence the design with $n = 4$ is more efficient than the design with $n = 2$.

5. When $p = 7$ and if six replications are available we have three alternatives to choose from (i) repeat $n = 2$ thrice (ii) repeat $n = 3$ twice and (iii) use $n = 6$ once. The number of accuracies in the three cases will be 4, 7 and 3 respectively. The efficiency factor in the first case is $1435/1879 = 0.764$ and in the third case $41615/49139 = 0.847$. The third alternative will give the most efficient design.

6. We may finally sum up this section by saying that the n-ple rectangular lattice is a p.b.i.b. design having 4, 3 and 2 associate classes respectively when $n = 2(p \geq 4)$, $n = (p-1)$ and $n = p$ respectively. It was demonstrated by the author (1951) that the n-ple rectangular lattice is not a p.b.i.b. design when $n = 3(p \geq 5)$. Probably this is also the case for higher values of n up to $(p-2)$.

III. *Dual of the n-ple rectangular lattice.*

1. If treatments and blocks of the lattice:

$$v = p(p-1), \quad k = (p-1), \quad r = n, \quad b = np$$

are called blocks and treatments, we get the dual design:

$$v^* = np, \quad k^* = n, \quad r^* = (p-1), \quad b^* = p(p-1).$$

2. When $n = p$, the dual design is a p.b.i.b. design with two associate classes. The parameters are:

$$v^* = p^2, \quad k^* = p, \quad r^* = (p-1), \quad b^* = p(p-1)$$

$$\lambda_1 = 1 \qquad \lambda_2 = 0$$

$$n_1 = (p-1)^2 \qquad n_2 = 2(p-1)$$

$$p_{ik}^1 = \begin{Bmatrix} (p-2)^2 & 2(p-2) \\ 2(p-2) & 2 \end{Bmatrix}; \qquad p_{ik}^2 = \begin{Bmatrix} (p-1)(p-2) & (p-1) \\ (p-1) & (p-2) \end{Bmatrix}$$

This design is a square lattice in which the rows, columns and $(p - 3)$ orthogonal squares are used.

3. When $n < p$ the dual design is a p.b.i.b. design with three associate classes. The parameters are:

$$v^* = np, \quad k^* = n, \quad r^* = (p-1), \quad b^* = p(p-1)$$

$$\lambda_1 = 1 \qquad \lambda_2 = 0 \qquad \lambda_3 = 0$$

$$n_1 = (p-1)(n-1), \quad n_2 = (p-1), \quad n_3 = (n-1)$$

$$p_{ik}^1 = \begin{bmatrix} (p-2)(n-2) & (p-2) & (n-2) \\ (p-2) & 0 & 1 \\ (n-2) & 1 & 0 \end{bmatrix}$$

$$p_{ik}^2 = \begin{bmatrix} (p-2)(n-1) & 0 & (n-1) \\ 0 & (p-2) & 0 \\ (n-1) & 0 & 0 \end{bmatrix}$$

$$p_{ik}^3 = \begin{bmatrix} (p-1)(n-2) & (p-1) & 0 \\ (p-1) & 0 & 0 \\ 0 & 0 & (n-2) \end{bmatrix}$$

When $n = (p - 1)$, the dual design has identically the same values for all the parameters as the original lattice, considered in paragraph 3 of Section II.

4. The dual design is identical to the design of the type (6.5) given on p. 122 of Nair and Rao's (1948) paper if we replace s_1 , s_2 , λ_{11} , λ_{01} and λ_{10} of their notation by n, p, λ_1 , λ_2 and λ_3 respectively.

5. In virtue of the fact that the dual of the n-ple rectangular lattice $(2 \leq n \leq p)$ is a p.b.i.b. design it becomes easy to analyse this lattice by first estimating the block effects and their sum of squares. These quantities can then be used to estimate the treatment effects and their sum of squares.

In conclusion, it appears to the author that although the n-ple rectangular lattice does not satisfy the conditions of a p.b.i.b. design except when $n = 2$, $(p - 1)$ or p, the reason why this lattice can be analysed in a systematic way for all values of n is due to the fact that the dual design is a p.b.i.b. design for all values of n.

REFERENCES

Bose, R. C. and Nair, K. R. Partially balanced incomplete block designs. *Sankhya*, 4, 337–372, 1939.

Fisher, R. A. and Yates, F. *Statistical Tables*, Oliver and Boyd, Edinburgh, Third edition, 1948.

Harshbarger, B. Rectangular lattices. *Virginia Agricultural Exp. Sta., Memoir*, 1, 1947.

Harshbarger, B. Triple rectangular lattices. *Biometrics*, 5, 1–13, 1949.

Harshbarger, B. Near balance rectangular lattices. *The Virginia Journal of Science*, 2, 13–27, 1951.

Nair, K. R. and Rao, C. R. Confounding in asymmetrical factorial experiments. *J.R.S.S.*, Series B, 10, 109–131, 1948.

Nair, K. R. Rectangular lattices and partially balanced incomplete block designs. *Biometrics*, 7, 145–154, 1951.

CORRECTION TO

"THE ESTIMATION OF RESPONSE-TIME DISTRIBUTIONS II"

M. R. Sampford

In the *Appendix Table* to the above paper (*Biometrics*, Dec. 1952) nearly all the values of the function ζ were erroneously printed as negative. These values are numerically correct, but should, in fact, all be positive.

The author deeply regrets this error, and wishes to express his apologies for any inconvenience or misunderstandings it may have caused.

QUERIES

GEORGE W. SNEDECOR, *Editor*

98 QUERY: A series of six samples of milk powder were sent to 7 different laboratories. The laboratories were asked to rank the powders, after testing, in order of quality with 1 as highest quality and 6 as low.

THE RATINGS ARE AS FOLLOWS:

Laboratory	Powders					
	A	*B*	*C*	*D*	*E*	*F*
1	3	6	1	2	5	4
2	1	3	4	2	5	6
3	2	3	5	1	4	6
4	3	1	5	2	6	4
5	1.5	1.5	6	4	3	5
6	3	4	5	1	2	6
7	4	1	5	2	6	3
Sum	17.5	19.5	31	14	31	34

The question then—is there a statistically significant difference in these placings or could such an arrangement happen by chance alone? The basic assumption is, of course, that Lab 1 (or any other lab chosen) placed the powders in their "proper" order.

I know the comparison of two sets of judgements and the application of the formula:

$$\rho = 1 - \frac{6\Sigma\Delta^2}{N(N^2 - 1)}$$

but this does not take into consideration comparisons of more than two judgements. I have an idea that the difficulty lies in selecting a value of N but just how to correct it I do not know.

ANSWER: The method to be used depends upon the underlying distribution of the rankings. If the scale can be assumed to be linear and the population distribution normal, the rankings may be transformed to normal deviates by the use of Table XX given in the Fisher and Yates tables. Analysis of variance would then be the appropriate statistical method.

If you think that there is no such underlying distribution, an appropriate method is described in Kendall's "Advanced Theory of Statistics," Volume 1, page 410; also in his "Rank Correlation Methods." An easy method of performing the test is furnished by Friedman in "Annals of Mathematical Statistics," Volume 11, page 86 (1940).

In your experiment, the results based on the two assumptions are substantially the same; the null hypothesis of concordance is rejected with $P < 0.01$. If your data warrant the change to normal deviates, you may wish to take advantage of the flexibility of analysis of variance.

99 **QUERY:** This query concerns a 5-point rating scale given to the same subjects before and after an experience with the object of the rating. Since 5-point and similar scales are widely used in research in some fields of psychology and in various applications such as industry, I think that this problem is of rather general interest.

A 5-point rating scale of interest was given to a group of subjects before and after a trial experience with a particular method. The question is, can the obtained difference be regarded as coming from a population of such differences where the mean is greater than 0 (or from a population of such differences where the mean is positive, or from a population of such differences where the mean is negative)? A further question is of lesser interest: are the two sets of ratings correlated? Will you please tell me which method should be used to answer these questions?

Additional light will be thrown on the data if I mention several possibilities for analysis which have occurred to me.

1) Chi-square test of independence from an $r \times c$ table. This test would indicate however only whether the relative positions of the individuals were approximately the same, and not whether the mean changed.

2) Chi-square for a $2 \times r$ table. I believe however that this does not take account of the correlation; all the examples I have seen have been two independent distributions rather than two correlated distributions.

3) Yates' chi-square test of independence using a regression estimate of column means on rows or row means on columns, applicable where (as here) the classes of the two attributes are quantitative. Again, this

would test only correlation. (Biometrika, 1948, 35, 176–181: The analysis of contingency tables with groupings based on quantitative characters).

4) A *t*-test could be made, but the "before" distribution is not very normal and the two variances are not very similar; and isn't it necessary to have a more continuous variable than 5-point one?

5) Non-parametric test, such as Festinger's based on ranks. But with only 5-points, the ranks would consist largely of ties (multiple ties, such as of 17 subjects). (The significance of difference between means without reference to the frequency distribution. Psychometrika, 1946, 11, 97–105.)

6) A sign test of the differences. (Dixon and Mood, The Statistical Sign Test. J. Amer. Stat. Assn., 1946, 41, 557–566.)

The original data are enclosed. I shall be very appreciative of your advice on this.

ORIGINAL DATA (FREQUENCIES)

Before	After					
	not at all 1	a little 2	some 3	much 4	very much 5	Total
very much 5	2	0	3	6	11	22
much 4	0	2	8	10	0	20
some 3	3	5	17	3	2	30
a little 2	0	3	2	0	0	5
not at all 1	4	0	0	0	0	4
Total	9	10	30	19	13	81

ANSWER: Two decisions must be made before your questions can be answered. One concerns the shape of the population from which you are sampling and the other, the scale of measurement. A sample of 81 pairs may give some indication of the facts but cannot be definitive. For this reason, the decisions should always be made in advance of the sampling. Your query suggests that this was not done.

With no more than the sample to guide me, I would not hesitate to use the ordinary methods devised for paired samples from normal populations. The differences, "after" minus "before", are tested by *t*; or, alternatively, the confidence interval may be used to guide you to conclusions. The correlation is calculated in the usual way.

I would place a good deal of confidence in the t-test because it is little affected by anormality, especially in samples so large as 81. As for the correlation, its existence is obvious, but an estimate of it may have a larger confidence interval than that computed by the usual methods.

Your bivariate frequency table lends itself readily to the calculation of the statistics you wish. The sample means, standard deviations, and correlation coefficient are got directly by the methods described in any text. The sample standard deviation for the t-test is then obtained from the formula

$$s_{\bar{d}}^2 = s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 - 2rs_{\bar{x}_1}s_{\bar{x}_2}$$

With no knowledge of anormality in the population you are sampling and with no more indication of skewness than that furnished by your samples I see no advantage in resorting to contingency tables or other nonparametric methods. But if you think that your ratings have little scalar validity, yet do serve to distinguish increased interest from diminished, you could use the sign test or its chi-square equivalent,

$$\chi^2 = \frac{(a - b - 1)^2}{n}, \qquad \text{d.f.} = 1,$$

where a and b are the frequencies in the two categories, increasing interest and decreasing. In your experiment there are 45 ties. The usual method is to allot half of these to each category. Chi-square (with Yates' correction for continuity) is 5.44, $P = 0.02$.

It is interesting to compare this result with that of the t-test on which the null hypothesis is rejected with $P < 0.001$. This decreased hazard of rejection corresponds to the superiority of that design for which the experimenter provides a linear scale with normal distribution. If he does not (or cannot) make such provisions he sacrifices the additional information that might be obtained.

THE BIOMETRIC SOCIETY

British Region. The British Region held its annual meeting at the National Institute for Medical Research in London on December 17th, 1952. The following officers were elected for 1953: *Vice-President*, Frank Yates; *Treasurer*, A. R. G. Owen; *Secretary*, E. C. Fieller; and *Committee* (1953-55), J. M. Tanner and J. W. Trevan. The business meeting was followed by a demonstration by W. L. M. Perry of some of the methods of biological assay used in the National Institute for Medical Research with the participation of D. A. Long, R. C. Grey, J. H. Humphrey, J. W. Lightbown, A. Isaacs, A. A. de C. Sampaio and Miss H. M. Bruce.

Région Française. La séance de la Société Française de Biométrie (Région Française) le 25 Fevrier au Laboratoire de Zoologie de l'Ecole Normale Superieure à Paris a constitué la assemblée générale annuelle. Communications ont été présentées par J. Sutter et L. Tabah sur "Résultats du test mosaique de GILLE dans des fratries entières de deux enfants et dans des couples de jumeaux", et par G. Teissier sur "Estimation des paramètres de la droite d'allométrie".

Region pour la Belgique et le Congo Belge. At a general meeting in Brussels on December 3, 1952, our Belgian members organized the "Société Adolphe Quetelet", following the pattern of the French group. We extend hearty greetings to this youngest Region of the Biometric Society! Its officers are: *Vice-President*, Paul Spehl; *Secretary*, Leopold Martin; *Treasurer*, Claude Panier; *Committee* (1952-1953), E. P. Cordiez, D. Demeulemeester, P. P. Deneyer, P. O. Hubinont, R. Laurent, J. Lebrun, Miss A. Lenger and J. Semal.

ENAR. The Region held a joint meeting with the Biometrics Section of the American Statistical Association at the Palmer House in Chicago on December 27 to 29. At the opening session under the chairmanship of B. G. Greenberg, papers were presented by H. L. Lucas, R. L. Anderson and A. M. Dutton on the "Statistical Analysis of Biological Time Series". A concurrent session with P. M. Densen in the chair considered "New Developments in the Use of Sample Surveys to Determine Health Conditions" with papers by N. R. Deardorff and J. Cornfield, by T. D. Woolsey and by P. M. Densen, A. Ciocco and D. Horvitz. In a late afternoon session, G. W. Schmidt, G. Karreman and A. Rapoport presented papers concerning the "Mathematics of Biological and Social Phenomena", with N. Rashevsky as chairman.

Sessions on "Two Recent Developments in the Design and Analysis of Experiments" and "Determining Optimum Conditions" were held on Sunday, December 28. C. Daniel was the chairman and J. W. Tukey and W. J. Youden were the speakers in the first of these, and H. Hotelling the chairman and R. L. Anderson and F. Mosteller the speakers in the second. On Monday, December 29, there were three sessions. The first was a combined program on "Randomization Theory and Its Relation to Non-Parametric Methods" with J. W. Tukey as chairman. Papers were presented by L. E. Moses, M. E. Terry, W. J. Dixon and O. Kempthorne. The first afternoon session considered "Problems and Activities of Experiment Station Statisticians", with H. W. Norton in the chair, papers by J. G. Darroch and C. E. Marshall, and discussions by H. Tucker, V. L. Anderson, P. G. Homeyer and R. J. Monroe. The closing session on "Determination of Means and Regression Coefficients" was under the chairmanship of F. S. Acton and featured papers by D. Wallace and R. F. Link.

WNAR, Regional Meeting, Stanford University, June 19-20, 1953. The joint regional meetings of The Biometric Society (WNAR) and of the Institute of Mathematical Statistics have been scheduled for Stanford University on June 19 and 20, 1953. The Biometric Society has sessions scheduled on the general topics of bioassay and population enumeration. One of the sessions of the Institute will be devoted to non-parametric inference. There will also be sessions for contributed papers. Abstracts of contributed papers to be presented at the meeting should be mailed to Assistant Secretary Rosedith Sitgreaves, Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, California, by May 28, 1953. All requests for accommodations should be forwarded to the Assistant Secretary. The chairman of the program committees for the Society and the Institute is Professor A. Bowker, of Stanford University.

Activities of National Secretaries. Our Indian group has now been reactivated, where we are fortunate in having Dr. V. G. Panse, Indian Council of Agricultural Research, Mansingh Road, New Delhi, as National Secretary.

In the past few months our membership in Germany has grown markedly, in large part through the efforts of Professor Dr. Maria Pia Geppert, W. G. Kerckhoff Institut, Bad Nauheim, who has been appointed National Secretary, and of Professor Dr. Richard Prigge of the Paul-Ehrlich Institut, Frankfort a.M.

A Japanese group is being formed and will be reported in the next issue of BIOMETRICS.

General Election. Recent elections named the following general officers for 1953: *President,* G. Darmois; *Secretary-Treasurer,* C. I. Bliss; *Members of Council for* 1953-1955, H. C. Batson, L. L. Cavalli-Sforza, W. G. Cochran, Sir Ronald Fisher, L. Martin and J. W. Tukey. We are grateful to the following members, who completed their terms as ordinary members of Council in 1952: M. H. Belz, P. V. Sukhatme and E. B. Wilson.

Policy Statement on BIOMETRICS. In order to bring BIOMETRICS before a wider circle, the Council has approved the following statements of policy. Comments on these or other policies of the journal are invited from the membership. They may be sent either to the Secretary-Treasurer (Drawer 1106, New Haven 4, Connecticut, U.S.A.) or to the Editor of BIOMETRICS.

1. Many papers presented at meetings sponsored or cosponsored by The Biometric Society, whether regional, national or international, are appropriate for publication in BIOMETRICS. The authors of all such papers should be encouraged to submit them to the Editor for consideration, even though BIOMETRICS is not thereby committed to their publication. Especially when a meeting is sponsored jointly by the Society and some other organization, the interests of both BIOMETRICS and the journal of the cosponsoring society should be considered. For instance, the methodology could appear in a section in BIOMETRICS on "Applications" and the interpretation of the data in the journal of the applied field with suitable cross-references. Separate articles with cross-references would give emphasis to both the interpretation and methodology, increase the demand and interest in BIOMETRICS and call the attention of readers of BIOMETRICS to the other journals.

2. At present, there is incomplete coverage in the publication of abstracts from meetings of the Society. Some regions have sent abstracts regularly and others have sent very few, if any. Abstracts of papers should be sent to the general secretary's office by the regional secretary or by some designated representative within a few days after each meeting.

3. Papers are sometimes submitted which on their merits would be approved for publication by the referees, except that the tables or illustrations are both too extensive and of too limited interest to justify this use of our limited funds. Sometimes, organizations are willing to pay the cost of printing such tabular material in order to obtain publication in BIOMETRICS. As a trial policy such papers may now be accepted with a footnote that the expense of publishing tabular or other material has been met by the organization sponsoring the research.

NEWS AND NOTES

In a recent letter, **Helen Turner,** McMaster Laboratory, Parramatta Road, Glebe, N.S.W., Australia, writes:

"In Sydney the Statistical Society of N.S.W. has gone smoothly on its way, holding monthly meetings. **Dr. A. H. Pollard** was President for a second year, and has just been replaced by **Dr. H. O. Lancaster,** who has been an energetic secretary for the last few years.

"The Biometric Society has continued to meet in Melbourne and Sydney. The Sydney branch now meets independently of the Statistical Society, and though numbers are sometimes small, some lively discussions have been held. The Society's main activity for the year was the inauguration of a Biometrics Session at the meetings of the Australian and New Zealand Association for the Advancement of Science, which was held in Sydney last August. This was a half-day session, run as part of Section D (Zoology), at the invitation of Section D. It was run jointly as an ANZAAS session and as the Biennial Meeting of the Australasian Region of The Biometric Society. Response from general ANZAAS members was good, and we hope to make it a permanent feature of the programme.

"Another innovation at ANZAAS was a pre-conference session on genetics, organised by **Dr. Rendel** (Animal Genetics Section, C.S.I.R.O.) and lasting two days. The first day was devoted to plant work, the second to animals. The animal session was run jointly with the newly-formed Australian Society for Animal Production, branches of which are meeting in Brisbane, Sydney and Melbourne. Genetics is booming in Australia—**Professor Catcheside** from Adelaide, and **Dr. Frankel,** the new Chief of the Division of Plant Industry C.S.I.R.O. were both at this meeting. A committee was appointed to organise a Genetics Society in Australia.

"Another event on the statistical front was the arrival in Canberra early this year of the newly-appointed Professor of Statistics at the Australian National University—**Professor P. A. Moran,** a Sydney graduate who has been some years in Oxford. I regret not being able to report the arrival of a lecturer in the Faculty of Agriculture at Sydney, to replace **Dr. D. B. Duncan.** This post is unfortunately still vacant, the lecturing being done temporarily by **Miss M. McKevett** (C.S.I.R.O.). Melbourne University advertised a full Chair of Statistics during the year.

"In cataloguing the expansion of statistical work in Australia I must unfortunately include a report of the death of one of our foremost

pioneers, **Mrs. Calvert,** who was formerly **Miss F. E. Allan.** She was the first Australian to go overseas specifically to study biometry, being sent by C.S.I.R.O. (then C.S.I.R.) to work at Rothamsted in the late 1920's. She worked as consulting statistician to C.S.I.R. until her marriage in 1940 and since then had lectured at the Forestry School at Canberra and worked part-time with the Bureau of Census and Statistics. She had suffered from high blood pressure for many years, and died quite suddenly last August. Her husband, **Dr. Calvert**, is still with C.S.I.R.O.; there is one son, Allan, now aged about ten.

"Events for the future in Sydney include a series of seminars on the statistics of populations; these are being organised for next year by Dr. Lancaster.

"My own work during the year has been a consolidation and confirmation of results from the previous year's work on sheep-breeding. I've visited our sheep breeding trials at Gilruth Plains, Cunnamulla twice—once to help in standardisation of observations in February (observer differences need watching in large-scale trials scattered over the countryside) and once at mating when final selections were in progress. We introduced one innovation this year in our selection trials—culling of ewes was on paper only, so that we may have a chance to check on our selection methods.

"Those of you who have been to Gilruth Plains may be interested to hear that 8,000 acres (about 25% of the total) were burnt out by bushfires early in the year. No stock or buildings were lost, but a considerable amount of fencing went. There have been press reports of another fire last week, but so far we've had no official details. Lambing is in progress, so we hope no damage was done.

"I've had the good fortune to be analysing an excellent set of sheep-breeding records kept by **Mr. Euston Young,** of Noondoo, Dirranbandi, Queensland. Mr. Young has been selecting his rams on progeny testing, with measurements similar to ours, except that density (fibres per unit area) has not been measured. I visited Noondoo in April and collected fleece samples on which all characters, including density, have just been measured. Indications are that we will get the same results as with our own animals—namely, that variations in density are more important than variations in the other components of clean wool weight.

"Animal husbandry-men will be interested in another activity in which we have participated this year, namely the training of graduate sheep and wool extension officers for State departments of Agriculture. Three officers were in this initial scheme, representing two states. They are graduates in agriculture or veterinary science, and have been spending varying periods during the year with one or another of the folk

actively engaged on sheep and wool research. This personal-contact method is a further attempt to bridge the gap between research and practice.

"Other items of interest are: **Professor C. E. Emmens** (Australasian Region Vice-President for the last two years) was recently appointed part-time Officer-in-Charge of C.S.I.R.O.'s new Sheep Biology Laboratory, at present being built near Sydney, and **Dr. Mildred Barnard's** husband, **S. Prentice** has for some time been Professor of Electrical Engineering at the University of Queensland, where Mildred and the four children joined him last year, after great difficulty in finding a house."

Summer Sessions at Berkeley, California

This year's summer program at the Statistical Laboratory of the University of California, Berkeley, California, consists of two sessions: June 22–August 1 and August 3–September 12. The faculty of the summer sessions will include Professor David Kendall of Magdalen College, Oxford University; Professor J. Neyman, Dr. T. A. Jeeves, and Mr. A. Shapiro of the Statistical Laboratory, University of California.

The program includes two of the usual undergraduate courses in each session. In addition Professor Kendall will give a new course in the first session, "Stochastic processes associated with population growth and with the theory of queues." This course is designed to acquaint students with the probabilistic treatment of growth of populations subject to birth, death, immigration, mutation and aging. Professor Neyman will be available for consultations on work leading to higher degrees.

The Southern Regional Conference on Statistics

On Oct. 3–4, 1952, the Southern Regional Education Board held a conference on statistics during which time, five groups considered the following topics: (1) additional training for existing personnel in statistics, (2) coordination of curricula in statistics, (3) statistical consulting services, (4) contact and cooperative research in statistics and (5) the basis for a regional program in statistics. The findings of the groups on these phases of statistical instruction, research and service in Southern colleges and universities as approved by the conference have been incorporated into a pamphlet recently released by the Southern Regional Education Board, 830 West Peachtree Street, N.W., Atlanta, Georgia.